Identifying NOx Sensor Failure for Predictive Maintenance of Diesel Engines using Explainable AI

Thomas Mckinley[1], Meghana Somwanshi[2], Devawrat Bhave[3], and Sandeep Verma[4]

[1,2,3,4] *Cummins Inc.500 Jackson St Columbus, IN 47201U.S.A.*

thomas.l.mckinley@cummins.com
meghana.somwanshi@cummins.com
devawrat.bhave@cummins.com
sandeep.verma@cummins.com

ABSTRACT

The automotive industry is being transformed by the application of artificial intelligence and big data analysis. In particular, predictive analytics is becoming a powerful tool for anticipating component failure. This key area of research provides automotive industry manufacturers with lower warranty expenses and incremental service parts revenue while rewarding customers with higher uptime. These benefits are particularly important for commercial vehicles operations such as bus and truck fleets, since analytics led predictive maintenance can prevent inconvenient and costly interruptions of vehicle mission. Accurate prediction models for component failures are especially challenging. This paper describes one such effort for failure prediction of transit bus NOx (Nitrogen Oxides) sensors.

Stringent emissions regulations have made diesel exhaust aftertreatment systems mandatory in nearly all global markets, and NOx sensors play a critical role in control and diagnostic algorithms used by these systems. NOx is measured before and after the SCR (Selective Catalytic Reduction) system with two different components namely Engine out (EO) NOx and System out (SO) NOx. Due to differences in operating conditions and failure rates these two components were studied separately. The results of different Machine Learning algorithms were obtained and compared to get the optimal predictions. Moreover, early life and late life failures were also studied separately to differentiate between random and wear-out failure modes. Highlights of the paper are: - the data collection process, feature engineering and feature selection process, as well as explainable AI (Artificial Intelligence) built on top of the machine learning model. Efforts were also taken to keep the approach generic and not become too component specific so that it can easily be replicated for predicting other failures on other product lines or of different components.

1. INTRODUCTION

Commercial vehicles are rapidly transforming from reactive repair to proactive maintenance with the help of Artificial Intelligence (AI) and high-performance computing cloud systems. Timely maintenance ensures reliable vehicle operation and avoids sudden breakdowns and interruption of vehicle mission. Moreover, significant financial benefits to the customer and manufacturer can be realized through reductions in repair and downtime costs.

Realization of these benefits, however, requires accurate failure predictions with sufficient advance notification, Over the last decade, with increasing applications/importance of predictive analytics, initial attempts have been documented in the literature. The Prognostics and Health Management Society (PHM Society) enables the advancement of these methods through shared research and application of PHM as an engineering discipline. Within this body of work, two different approaches have emerged: (a) physics-based models; and (b) data-driven models. This paper presents a data-driven approach for predicting failure of one of the crucial components related to aftertreatment which is NOx sensor of transit bus. (physics based (Daigle & Goebel, 2011) (Bolander, Qiu, Eklund, Hindle, & Rosenfeld, 2009) data-driven approach (Si, Wang, Hu, & Zhou, 2011) (Gurung, Lindgren, and Bostrom, 2017).

Although AI has been considered a black-box approach for relating key driving parameters to a predicted value of interest (in this case failure probability), there is immense

value in breaking the black-box paradigm. The resulting model is not simply the key to unlocking the financial benefits mentioned above. Rather, the sensitivity and importance of each parameter or feature to the predicted value is its own valuable and rich source of knowledge. Explainable AI and Shapley plots are shown in this paper to be vital tools for attaining this insight, which can be leveraged in several ways. First, it allows model results to be compared to the experience of subject matter experts. This can lead to the discovery of new insights as well as to new and improved models. Second, these insights can inform how vehicle usage may be adjusted to delay failure. Third, the identification of factors influencing failure can be used to improve both the product and the product validation process.

2. PROBLEM STATEMENT

Due to their durability and fuel economy, diesel engines are the power-plant of choice for commercial vehicles such as trucks and buses. However, high temperature zones during the combustion process lead to the formation of NO and NO2. In response to health effects studies, permissible levels of these pollutants on new engines has been reduced by roughly two orders of magnitude. Aftertreatment systems downstream of the engine play a key role in emissions reduction, and typically lower emissions levels by factors of 10 to 20. The most common NOX aftertreatment system is ammonia-based selective catalytic reduction (SCR) in which a urea-water solution, called diesel exhaust fluid (DEF), is injected into the exhaust upstream of a catalyst. The urea decomposes into ammonia (NH3) which is adsorbed on the catalyst, where it reacts with NOX and is converted to harmless nitrogen (N2).

NOX sensors are installed upstream and downstream of the SCR system to measure the conversion efficiency of NOX to N2. This measured value is a key input to control algorithms that command the injection rate of urea-water solution at each instant of time, as engine conditions dynamically vary in response to the driver's commands of throttle position and gearing. Sensor errors lead to DEF injection rate being higher or lower than intended. A higher than intended rate can lead to ammonia emissions, which irritate the lungs. A lower than intended rate can lead to excessive NOX emissions and their subsequent health effects. As a preventative measure, regulatory agencies require on-board diagnostics that nearly continuously monitor conversion efficiency. Given the central role of NOX sensors in conversion efficiency estimation and SCR system control, separate on-board diagnostics that monitor proper operation of both sensors are also required.

Should these diagnostics detect a problem with either sensor, the driver is alerted through a lamp on the dashboard and fault codes are recorded in the engine's electronic control module. To encourage replacement of the faulty sensor, some regulators require than the engine's power be reduced. The bus or truck journey is terminated, in some cases nearly immediately. Costs associated with unexpected repairs, lost revenue due to interrupted operation, and the need for additional staff and vehicles to serve (in some cases for several days) in place of the failed vehicle all are powerful motivations to avoid these unexpected repairs.

If instead the failure can be anticipated, and the sensor is replaced as part of a pre-planned maintenance event, the added costs and pain of the corresponding repair can be greatly reduced. This is a challenging problem since the sensor has multiple failure modes. For example, one failure mode is cracking caused by liquid water on the sensor during its warm-up cycle. This 'random event' failure can be sensitive to ambient humidity and exhaust gas temperature shortly after the engine is started. Because it is a random event failure, it is very prevalent in early-life failures. Moreover, there are also several late-life failure modes due to degradation of the sensor body heater and build-up of contaminants on the sensor probe. Sensor life for 'wear out' failures would be expected to be functions of the age of the sensor, the duty cycle to which the sensor is exposed, and noise factors such as ambient conditions and installation differences between different truck and bus models.

It is therefore not surprising that sensor failure rates can vary by a factor of two or more between fleets. As such, it is neither practical or desirable to choose a common maintenance interval. Rather, the ability to determine sensor failure probability for a given fleet, and to use that to choose the best maintenance interval, would be highly preferable. This determination would be best accomplished as a prediction rather than an analysis of failures already experienced by a particular fleet, so that time and pain associated with determining the maintenance interval can be minimized. Moreover, the predictive model needs to be very efficient, since there are hundreds of major bus and truck fleets in the United States alone. These objectives, combined with the complexity and multiplicity of failure modes, suggest this problem as an ideal opportunity for machine learning.

In this paper explains efforts around identifying the NOx sensor failure within warranty. For engines which are out of warranty, estimating engines Remaining Useful Life (RUL) would be appropriate (Si, Wang, Hu, & Zhou, 2011). We intend to publish effort around RUL separately followed by this paper.

3. DATA COLLECTION & DATA PREPARATION

In this study, three main sources of data are used. They are: (a) reliability data; (b) Engine snapshot data; and (c) ambient condition data. These data sources and their original

purposes are described below, followed by the specific features used by the predictive model.

Reliability data was originally collected for the purpose of warranty claim filing and payment (Lawless, J., Hu, J., & Cao, J. (1995). It includes a description of the engine (e.g. type, date of manufacture, serial number, date entered in service, vehicle identification number), the name of the owner, and all warranty claims for each engine (e.g. date of repair, odometer reading at repair, type or repair, parts replaced, location of repair shop, service technician narrative). The utility of this data is that it identifies which particular engines have experienced either an engine out or a system out NOX sensor failure, and when these failures occurred. It also identifies which engines are owned by a particular bus or truck fleet.

Engine snapshot data was collected from the engine's electronic control module (ECM) during repair events, and its original purpose was to assist the service technician in troubleshooting the problem with the engine. To that end, it includes a list of active and inactive fault codes and the number of times they were triggered as well as the number of times the control system entered 'engine protection' mode (e.g. high coolant temperature, high intake air temperature). Snapshot content also includes data used by the manufacturer to associate failures with duty cycle effects. Some of the duty cycle parameters available include histograms of engine speed – engine load combinations, metrics of vehicle speed, and similar other ones. Availability of these snapshot is subject to connecting the ECM device at authorized workshop.

Ambient condition data was originally collected to provide an almanac of weather data for use by the public. The particular data source used in this study was taken from the National Oceanic and Atmospheric Administration (NOAA) which compiles meteorological data (e.g. ambient temperature, ambient pressure, relative humidity) for numerous sites across the United States for each hour of each day and has done so for several years.

A novel approach was used to associate ambient condition data to individual engines by combining it with reliability data. Specifically, the repair location was used to find the closest NOAA weather site, and hourly data was combined over the time span from the date entered into service until either the failure date or the latest ECM snapshot on record. Moreover, the scope of the analytics effort was restricted to transit buses, which are primarily operated locally. In this way, it was possible to introduce ambient condition features which were not directly measured on the vehicle (humidity and ambient pressure) but which were found to be important.

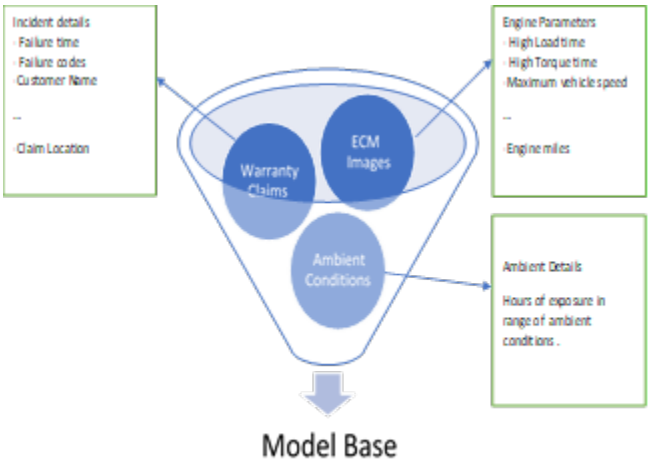

Figure 1. Three main sources of data used in the study

Features Related to Duty Cycle:

Duty cycle is expected to affect both random event failures as well as wear-out failures. Random event failures would be sensitive to the number of key cycles (on-off events), which increase the potential for this failure to occur. Heater failures would be sensitive to time spent at low exhaust temperature, where the electrical current must be increased to maintain sensor temperature. Since exhaust temperature is not recorded in any of the data sources, the model must use surrogate values for it. This is mainly evinced as operation at light load and at idle. A partial list of duty cycle features includes:

Load – Light and Heavy load time for an engine
Torque time – High Torque and Low torque time
Vehicle Speed – Ranges from 0 to max to 120 miles
Engine Speed – Speed of the engine
Coolant temperature
Key switch Cycle
Manifold Air temperature
Distance covered in top gear
Power – Averaged out of that engine
Coast Distance converted (in Km.)
Idle fuel used
DEF used in liters

Features Related to Age or Usage:

Age is expected to affect random event failures since the number of chances for a random failure would be expected to increase over time. Moreover, age should also be important for wear-out failures as contamination and damage accumulate over time. A partial list of age or usage features includes:

Engine Age – Calculated from Engine in service till latest date for non-failures and till failure date for failures

Engine Miles & Hours per month – Utilization parameters

Features Related to Ambient Conditions:

Ambient conditions have a complex impact of failure modes in that they can affect the amount of water in the exhaust gas and therefore the chance of a random cracking failure. Ambient temperature and pressure also have impacts on exhaust temperatures. Furthermore, extreme cold or heat can lead to increased idle time and / or idle engine speed for either warming up or cooling down the passenger compartment (Aliramezani, M., Ebrahimi, K., Koch, C. R., & Hayes, R. E. (2017)).

As mentioned above, ambient condition variables for a specific engine are taken from a nearby location and they are cumulative in nature over time. Account must also be made for variation, and in this study we use histograms with three bins. This same approach is used for ambient temperature, ambient pressure, and humidity. The limits for the three bins were iterated to find break points that were indicative of failure probability.

For example, if a variable considered is ambient temperature, it is represented in 3 bins where each bin would have a count for how long the truck operated under that particular temperature range. We call the exposure that engine had in that specific temperature range. The count of ambient temperature variable is always increasing for an ESN for every snapshot taken after the another. This means that different exposure hours for the same engine often is highly correlated. However, for our purposes, we will only select one datapoint per engine. Rather than choosing a random datapoint, we want the datapoint to be the most informative and closer to the NOx sensor failure. So, the cumulative exposure in a specific temperature bin for that particular engine will be the last snapshot taken before the breakdown occurred is considered. If for example for ambient temperature, the 1st bin would correspond to hours of operation in cold (warm) weather, 2nd bin in moderate weather and 3rd bin in warm weather. We also thought about representing these 3 bins as separate numeric attributes but increasing dimensionality had a negative impact on predictive performance. End goal for including ambient conditions was to analyze the impact of the condition on the failure. For example, if bin 1 gives a good separation for ambient temperature for engines with faulty and healthy NOx sensors; then the operation in cold weather can be considered to be a useful factor for predicting failure. We have used the repair location and the last snapshot date as the parameters to get the cumulative exposure for a particular engine. We would have preferred the actual failure location, but we did not have the data for the failure location and repair location was the closest that we depended on. For instances where the repair location was not available we have taken registration location for those specific engines. For engines where no datapoint related to its location was available, for such engine's location was imputed basis its co relation with the fleet owners. Rest of the engines where no data was available, such instances were dropped from the study. We have considered below 5 ambient variables. Each variable was divided into 3 bins. Temperature, Pressure, Humidity, Absolute Humidity & Dewpoint.

Transit bus engines from Untied States which are still within warranty are considered for this study. As one engine can have multiple snapshots with varying intervals, latest complete engine snapshot is selected for analysis. In particular, for failed engines, snapshot prior to NOx sensor failure is desirable. This adds predictive capability to NOx sensor failure model we intend to develop.

Features Related to System Interactions:

NOX sensors are a key component of the engine's aftertreatment and emission control systems. As such, their failure may be related to interactions with these rather complex systems. Engine history (i.e. various failure codes for warranty claims and fault codes recorded in the engine control module) over the last 3 months (90 days) are expected to provide important clues to these effects. One specific finding was that SONOx sensor failure occurred quite frequently within 2-3 months after EONOx sensor failure. In addition, active and in-active fault codes triggered also had some association with NOx sensor failure and were added to the model as features.

Exploratory Analysis and Hypothesis Testing:

Once the data sources were identified, exploratory analysis was conducted to gain insights as to the few factors most informative of NOx sensor failures. This effort revealed interesting points of immediate help to project stake holders. Univariate as well as multivariate analysis were employed to identify new features to be engineered. During this phase of the project there was strong collaboration between data scientists and engineers, sensor experts, and reliability engineers.

Some key findings are as follows. First, near 4600 engine hours, there is a transition to a much higher failure rate. This can be attributed to the onset of late-life wear-out failure modes. We also found that most of the time the EONOx sensor fails before the SONOx sensor fails. Using feature engineering concepts, we created a new feature that indicates prior EONOX sensor failure for the SONOX sensor. In addition, significant trends could be driven by fleet or location differences, or by the number of units in each location. Maximum vehicle speed was one such parameter where lot of research was done, and we found out that this parameter is somewhat lower on units with failures. This may indicate operation exclusively in urban environments.

As the analysis was done separately for EONOX and SONOX sensor, we also found differing insights when the failure populations were compared. For example, the average heater duty cycle (therefore the heater power) is higher on the EONOX sensor than the SONOX sensor and this contributes to its shorter life due to heater degradation. Similar studies were done on 40 parameters.

- What Features – Snapshot, fail code & fault code, reliability, ambient (no histogram, just high medium and low bucketing basis distribution), etc.

4. MODELING APPROACH

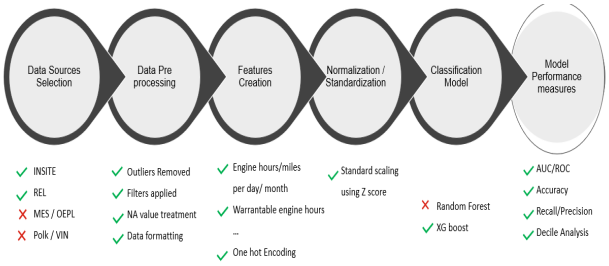

Figure 2. Model Building Process Followed

Machine Learning has become popular in recent times because of its accuracy and learning dimensionality. Performance of a machine learning algorithm purely depends upon the training data given to the algorithm. When it comes to multivariate analysis as a binary classification problem, where in this case the target variable is a failure or non-failure, XG boost gives a better performance over multiple machine learning method (Zhou, & McArdle, (2015), Ishwaran, Kogalur, Blackstone, & Lauer, (2008)).

XGBoost is a highly optimized implementation of a gradient boosted tree algorithm (Ridgeway, G. (2007), Ye, J., Chow, J. H., Chen, J., & Zheng, Z. (2009, November)), which involves training a sequence of increasingly predictive decision trees. At each "boosting" step, a new model is trained by fitting to the residuals from the previous step. The algorithm has proven its mettle in terms of speed and performance since its introduction in 2014. (XGBoost: A Scalable Tree Boosting System Tianqi Chen, University of Washington Carlos Guestrin, University of Washington)

In XGBoost, there is a good tradeoff between high performance and efficiency. As it can handle missing values in data, it eliminates the need for imputation or removal of entire sets of rows/columns. Being a tree ensemble-based algorithm (Johansson, Boström, & Löfström(2013, December)), it is interpretable but sometimes optimization can be tricky due to the presence of many hyper-parameters that need to be tuned.

In this study, we had some important variables with 20% missing values. Removing them from the data would have been undesirable. Instead, XGBoost naturally admits sparse features for inputs by automatically learning best missing values depending on training loss and handles different types of sparsity patterns in the data more efficiently. It also employs the distributed weighted Quantile Sketch algorithm to effectively find the optimal split points among weighted dataset as well as it is designed to make efficient use of hardware resources. As the solution was being developed on a Data Science Virtual Machine, efficient use of hardware was one of the key points in using this algorithm.

We also needed better interpretability which can be provided by a tree-based model (Shrikumar, Greenside, & Kundaje. (2017, August)). Due to these considerations we decided to use the XGboost technique to build the classification model. In order to tune the hyperparameters, we have used the cross-validation function from XGBoost

5. MODEL RESULTS

While building the model the data was divided into train (70%) and test (30%) and the results for the training set and testing set were examined separately.

We have used multiple techniques/metrics to analyze the results. Listing two important and effective metrics below that we get from a sci-kit learn package in python (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel & Vanderplas (2011).).

Confusion Matrix:

A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data

ROC (Receiver Operating Characteristic):

The ROC chart provides a means of comparison between classification models. The ROC chart shows false positive rate (1-specificity) on X-axis, the probability of target=1 when its true value is 0, against true positive rate (sensitivity) on Y-axis, the probability of target=1 when its true value is 1. Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases.

Here are the results for Engine Out NOX sensor:

Table 1. Confusion Matrix for EONOX failure predictions on train data

Train Data Confusion Matrix - EONOX			
	Predicted Label		
		NOX Not Failed	NOX Failed
True Label	NOX Not Failed	77%	2%
	NOX Failed	6%	15%

Table 2. Train Data Model Performance Metrics

Train Data - model performance metrics	
Precision	0.92
Recall	0.86
Accuracy	0.91
ROC	0.84

Table 3. Confusion Matrix for EONOX failure predictions on test

Test Data Confusion Matrix - EONOX			
	Predicted Label		
		NOX Not Failed	NOX Failed
True Label	NOX Not F	76%	3%
	NOX Failed	9%	12%

Table 4. Test Data Model Performance Metrics

Test Data - model performance metrics	
Precision	0.88
Recall	0.82
Accuracy	0.88
ROC	0.79

We calculated accuracy, precision, recall and specificity using the confusion matrix listed above. We can see that the results for train and test are nearly same. The mis classification rate for train and test is also nearly same suggesting that the model is not an overfit.

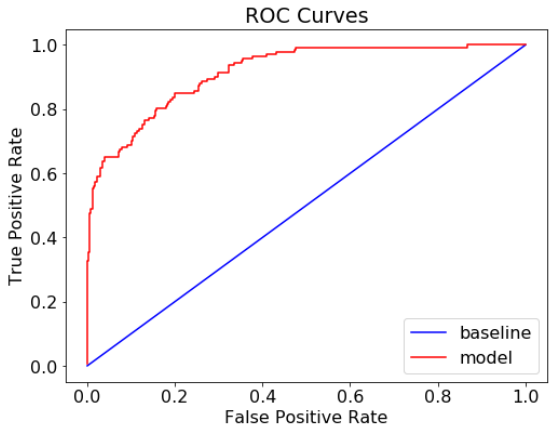

Figure 3. ROC Curve - to indicate model performance

Deciling, Gain and lift charts

Classification model's performance can be measured in number of ways. We selected a visual way to get an idea of how well a model is fitting the data by taking a look at the decile analysis

Decile analysis is a visualization technique that divides the data into 10 equal parts. In our case this division is performed as follows:

1. NOx sensor failure probability model is developed on training data
2. Holdout sample or test data is scored with probability between 0 and 1 using this model
3. The engines are then sorted in descending order of failure probability and then divided into 10 equal groups called deciles. Top decile has engines with high probability of failure and bottom decile with lowest probability of failure.

Decile proves to be a very efficient technique, when targeted intervention or predictive maintenance are done with budget constraint. It helps to select the engines which are highly likely to fail. i.e. by targeting say, top 3 deciles we can cover maximum failures. The deciles and their actual response rates are tabularized Return on investment can also be useful to decide the cut-off for targeting engines. This can be done by associating costs of recalling and maintenance for engines and saving with warranty and / or downtime cost one could avoid by recalling these engines which are highly likely to fail in near future.

Table 5 . Decile Analysis of EONOX failures

Decile	Non-Fail	Fail	Total	% Failure	Model	No Model	Lift
1	1%	45%	10%	99.1%	46%	10%	4.6
2	5%	29%	10%	63.7%	75%	20%	3.8
3	8%	12%	10%	24.9%	87%	30%	2.9
4	10%	7%	10%	14.8%	93%	40%	2.3
5	11%	3%	10%	8.3%	97%	50%	1.9
6	12%	2%	10%	3.7%	99%	60%	1.7
7	13%	1%	10%	1.2%	100%	70%	1.4
8	13%	1%	10%	0.3%	100%	80%	1.2
9	13%	0%	10%	0.6%	100%	90%	1.1
10	13%	0	10%	0.0%	100%	100%	1.0

Table is the decile table for our model results.

The decile analysis is suggesting that the model is "binning" the elements correctly from most likely to fail to least likely to fail. Our model exhibiting a good staircase decile analysis and can be considered to score the entire population. We also derived two metrics Gain and Lift to showcase the confidence in our model.

Gain or lift is a measure of the effectiveness of a classification model calculated as the ratio between the results obtained with and without the model. Gain and lift charts are visual aids for evaluating performance of classification models. However, in contrast to the confusion matrix that evaluates models on the whole population gain or lift chart evaluates model performance in a portion of the population. The lift chart shows how much more likely we are to get more failures in the top 3 deciles according to the model instead of a random selection of a sample. Below lift chart show that model performance is 5 times better than random selection.

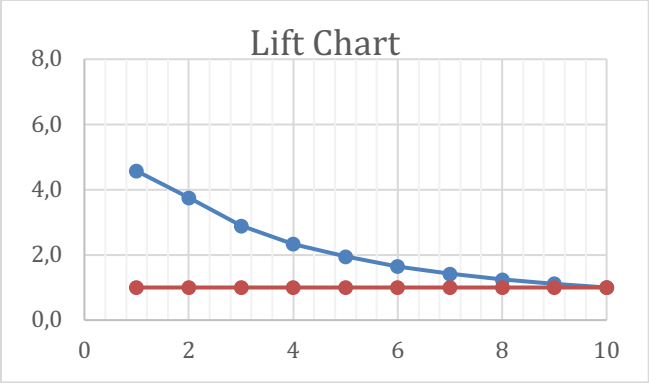

Figure 4. Lift Chart showing model performance is 5 times better than random selection

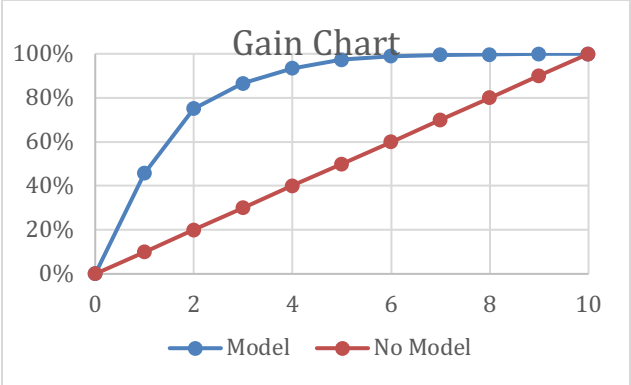

Figure 5. Gain chart showing amount of gain model gives relative to random model

Explainable AI and SHAP

As machine learning becomes our go to method to solve business critical problems, interpretability of a machine learning model becomes an important consideration. For models like XGBoost, where we usually have a robust accuracy metric, it becomes necessary to make the model more interpretable to identify factors that contribute directly towards the metric.

When the model is about enter production the bare minimum expectation out of a machine learning model is to work as expected and produce transparent explanations and reasons for decisions that are made.

Explainable AI, also known as XAI, is an emerging field in machine learning that aims to address interpretability of machine learning models. This area inspects and tries to understand the steps and models involved in making decisions. XAI answers some important questions like:

- On what basis did the algorithm decide to make a specific prediction?
- Does the algorithm give enough confidence in the decision?
- How can the errors be corrected?

Specifically, we used the recently introduced SHAP package, which utilizes techniques from Game Theory, to quantify feature importance at a highly granular level. SHAP (Shapley Additive exPlanations) provides a unified solution for all of our interpretability requirements. (Lundberg & Lee, 2017). We have used the term interpretability to indicate an extent to which a cause and effect can be observed within a system. We have used the term explainability when we wanted to indicate importance of feature values in relation to model prediction

Shapley is one of the XAI techniques that is based on Interaction based method for explanation. It has optimized functions for interpreting tree-based models and a model agnostic explainer function for interpreting any black-box model for which the predictions are known. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. the Shapley value for each variable (payout) is basically trying to find the correct weight such that the sum of all Shapley values is the difference between the predictions and average value of the model. In other words, Shapley values correspond to the contribution of each feature towards pushing the prediction away from the expected value.

SHAP is a measure of feature importance which is backed by theoretical guarantees of the following desirable interpretability objectives:

1) A consistent metric to assess feature importance.

2) Ability to visualize the marginal dependence of a given feature at a more granular level. (at unique engine level).

3) Ability to visualize interaction effects between chosen features at the level of individual observations.

The SHAP metric attributes feature importance locally (for each individual observation (Engine Unique number) in the training dataset) based on the Shapley value from Game theory. It is based on the underlying assumption of features to be playing a cooperative game where the payoff is the multivariate output (likelihood of being an issue), which then needs to be distributed fairly among the features. This approach has been shown to yield consistent results as well as provide insights into the failure dynamics with a high level of granularity, since it estimates feature importance at the level of each individual engine in the training data. For a given incident (observation), a positive SHAP value for a feature indicates that including that feature increases the likelihood of the incident being an issue, while a negative SHAP value indicates that the feature decreases the likelihood of being an issue.

The SHAP plots for our analysis are shown below. These plots give feature rank of 8 sensor values and its discriminatory power to determine the probability of target variable. Plot 1 Figure 6 suggests how is the relation of the features with the target. Red color means the feature has positive impact and blue color means the feature has negative impact. Only feature 0 has negative impact over NOx sensor failure but the other features have negative impact. Features – maximum vehicle speed, DEF used in liters, Low speed Medium & High Torque Time, Idle fuel used have negative impact on the target variable i.e. higher value increases failure rate.

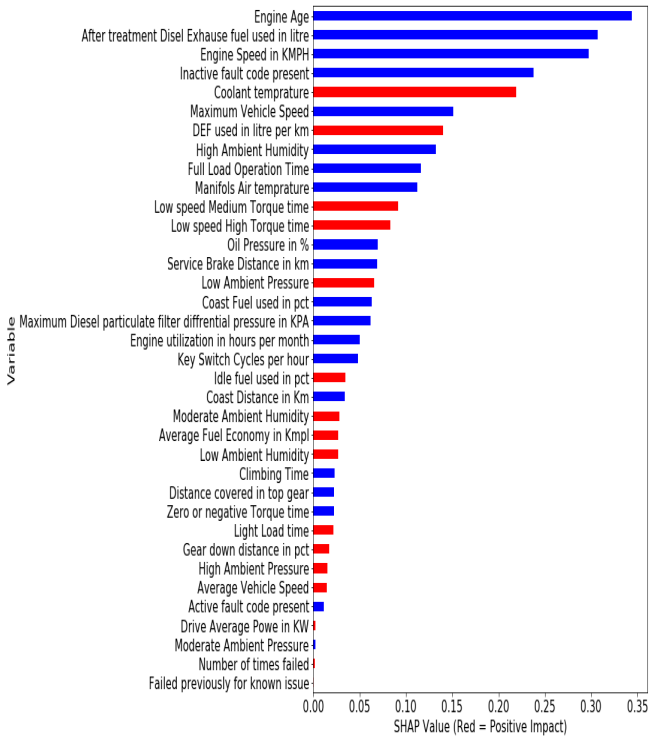

Figure 6. SHAP - feature importance chart for EONOX model with negative and positive effect of every explanatory variable on NOx Sensor Failure

The above plot (Figure 6) gives the feature importance that is different from the feature importance given by the XGBoost plot. For XGBoost the feature importance is defined by the weight / gain or coverage parameter. But, feature importance in SHAP has two benefits.

First one is global interpretability — the collective SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable. This is like the variable importance plot, but it can show the positive or negative relationship for each variable with the target

Second benefit is local interpretability — each observation gets its own set of SHAP values (see the individual SHAP value plot below). This greatly increases its transparency. We can explain why a case receives its prediction and the contributions of the predictors. Traditional variable importance algorithms only show the results across the entire population but not on each individual case. The local interpretability enables us to pinpoint and contrast the impacts of the factors

The reason we thought that SHAP was more appropriate in this analysis w.r.t. other competitors such as LIME was because we also wanted to see if how the explanatory variables were affecting the failures. For example, we know

that engine speed is an important variable to classify failures and non-failures. But the interest was to see if high engine speed was the causing failures. We get this kind of association score in SHAP. SHAP value guarantees a fair distribution of contribution for each of the variables and is optimized for tree-based prediction algorithms where things run very fast and the output is accurate and reliable. In contrast, LIME does not guarantee to perfectly distribute the effects. It builds sparse linear models around each prediction to explain how the black box model works in that local vicinity. LIME seems to be subset of SHAP. There are two important benefits which made us use SHAP instead of any other library -

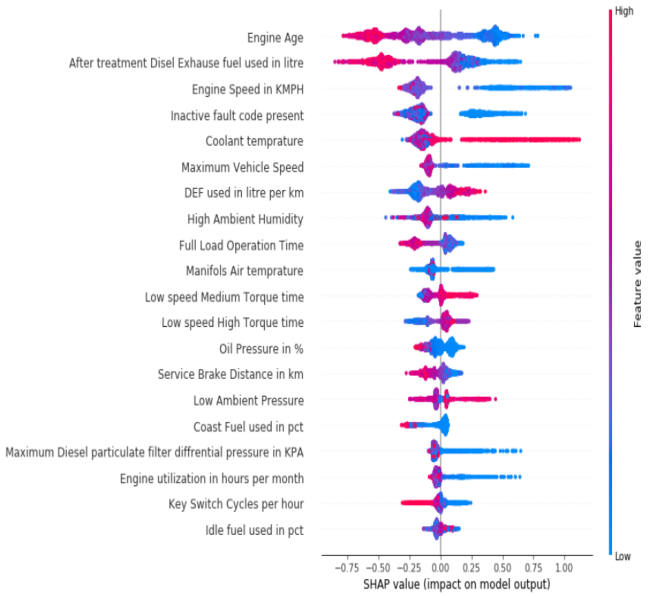

Figure 7. SHAP plot quantifying impact of every explanatory variable on model output

According to Figure 7, we can easily interpret the plot and it seems evident that features Engine Age, Engine Speed, Coolant Temperature, Maximum vehicle speed have greater discriminatory power which is useful for our predictions. Moreover, the plot also suggests the direction of impact. Red color represent high value of feature and X-axis represent the impact of feature on NOx failure(outcome). For example, higher the Engine Age (DEF usage) lower the risk of NOx failure. On the other side, high Coolant temperature has high risk on NOx failure. Feature at the bottom of the chart has relatively less impact on NOx failure as the SHAP impact is close to 0.

The results mentioned above are for Engine Out NOx sensor failure. We also followed similar approach to get feature importance for System Out NOx sensor as shown in Figure 8.

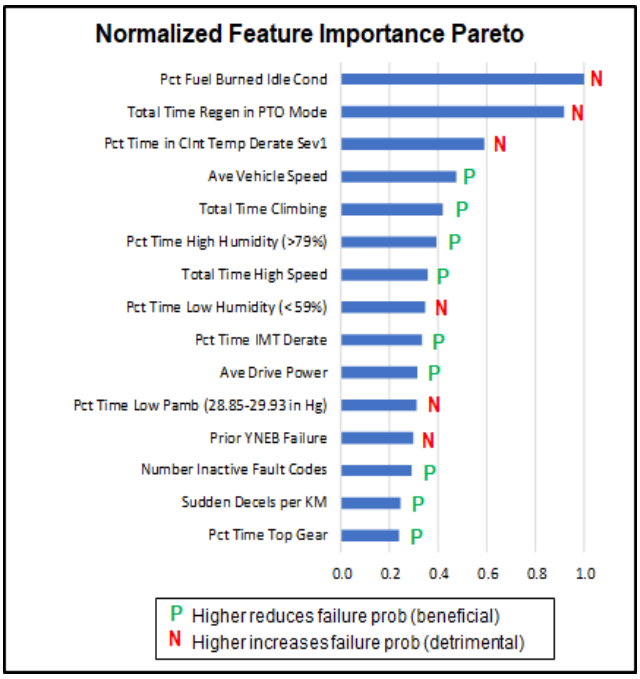

Figure 8. Regular feature importance chart with negative and positive effect on SONOX Failure

6. CONCLUSION

The objective of this study was to demonstrate that EONOX sensor or SONOX sensor failure during the warranty period could be accurately predicted using available data. The results mentioned above show that this has been accomplished, setting the stage for preventive replacement of sensors using either unit specific intervention or fleet specific maintenance intervals.

This was a challenging task, since NOX sensors have multiple failure modes, some of which are due to random events. Keys to success included:

- Minimizing the influence of random failures, which comprise about 10% of all failures during the warranty period, by screening out failures occurring prior to 4600 engine hours. This allowed model training to focus on more predictable wear-out failure modes.
- Using exploratory analysis in conjunction with subject matter expertise to rapidly accelerate feature engineering and the screening of features to be included in the model.
- Utilizing public site weather data to include new, important features to the existing data set and combining it with repair location to associate weather data to specific engines.
- Including system interactions through the consideration of prior failures (warranty claim fail codes and fault codes).

- Employing the Xtra Gradient Boost machine learning technique to train the model.
- Sharing SHAP plots with project stakeholders to screen for model interpretability and rationality.

Next steps include preparing the model for production so that business benefits can be achieved and developing a remaining useful life model which can provide a more precise estimate of failure timing. The techniques developed and proven here will also be reapplied to other key engine components.

ACKNOWLEDGEMENT

We are especially appreciative of technical insight and guidance provided over the course of this project by Dr. Nilesh Powar from Cummins. Mr. Amit Pandey from Cummins is thanked for connecting us to prior work on NOX sensor prognostics. Mrs. Subhalakshmi Behera provided us key guidance on planning all the activities and helping us understand business requirements. Mr. Christopher Newman from Cummins provided valuable understandings of reliability data, fleet composition, and failure probability variation across different fleets. Mr. Virendra Parte provided us key insights regarding data extraction and storage. Mr. David Hall provided us necessary technical and engineering guidance on NOx sensor failure. Mrs. Meenu Tomar provided us valuable understanding of targeted intervention and hypothesis around NOx sensor failure All of these colleagues were key contributors to this effort. Finally, the authors thank Cummins, Inc for the opportunity to publish this work.

NOMENCLATURE

XGBoost - eXtreme Gradient Boosting
NOX / NOx- Nitrogen Oxide
EONOx – Engine Out NOx
SONOx – System Out NOx
SHAP - Shapley Additive exPlanations
XAI – Explainable AI
AI – Artificial Intelligence

REFERENCES

Gurung, R. B., Lindgren, T., & Boström, H. (2017). Predicting NOx sensor failure in heavy duty trucks using histogram-based random forests. *International Journal of Prognostics and Health Management*, 8(1), 1-14.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, *1*(1), 2007.

Ye, J., Chow, J. H., Chen, J., & Zheng, Z. (2009, November). Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 2061-2064).

Shrikumar, A., Greenside, P., & Kundaje, A. (2017, August). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3145-3153). JMLR. org.

Aliramezani, M., Ebrahimi, K., Koch, C. R., & Hayes, R. E. (2017). Investigating the effect of temperature on NOx sensor cross sensitivity to ammonia using a physics-based model.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).

Si, X. S., Wang, W., Hu, C. H., & Zhou, D. H. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1-14.

Zhou, Y., & McArdle, J. J. (2015). Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, *80*(3), 811-833.

Shwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841-860.

Johansson, U., Boström, H., & Löfström, T. (2013, December). Conformal prediction using decision trees. In *2013 IEEE 13th International Conference on Data Mining* (pp. 330-339). IEEE.

Lawless, J., Hu, J., & Cao, J. (1995). Methods for the estimation of failure distributions and rates from automobile warranty data. Lifetime Data Analysis, 1(3), 227-240.

Daigle, M. J., & Goebel, K. (2011). A model-based prognostics approach applied to pneumatic valves. *International Journal of Prognostics and Health Management*, *2*(2), 84-99.

Gazizulin, D., Klein, R., & Bortman, J. (2018). Physics Based Methodology for the Estimation of Bearings' Remaining Useful Life: Physics-Based Models, Diagnostic Methods and Experiments. In *Fourth European Conference of the PHM Society*.

BIOGRAPHIES

Thomas L. McKinley is the Corporate Director of Engineering Analytics at Cummins, Inc. He holds bachelor and master's degrees in mechanical engineering from Purdue University at West Lafayette and a doctoral degree in Mechanical Engineering from the University of Illinois at Urbana-Champaign. His multi-decade career has included engine combustion and emissions modeling, product development, problem solving, physics-based product validation methods, and engineering analysis of a wide range of diesel engine emissions control systems. He has authored 18 technical publications, including six journal articles. In recognition of this expertise, he has been elected as a Fellow in the Society of Automotive Engineers (SAE) and served as an associate editor for the SAE International Journal of Engines. Current research interests include prognostics for critical engine components.

Meghana Somwanshi, is Data Science professional with both master's and bachelor's degree in Statistics from University of Pune. She has over 14 years of experience in hospitality, telecom, and manufacturing domains. She has primarily worked on targeted interventions, lifetime value modeling, and incremental revenue optimization. Her area of interest includes working on analytical solutions for insightful data driven decisions and research interest in Machine learning and Artificial Intelligence.

Devawrat Bhave, born in India, is a Data Science professional working in Cummins and holds a Master degree in Data Science from Praxis Business School. He is working the field of data science for last 7 years. His main area of interest is the study of various machine learning algorithms and their implementations in various domains. His work experience includes working for domains like Life sciences, Finance and Manufacturing domains.

Sandeep Verma is an Engineering graduate and a Data Engineer having 13 years of Industrial exposure in various domains including Manufacturing. He has specialization in legacy modernization, Creating Data Archival platform and building IOT data pipelines. His area of interest is to explore various data centric, cost effective cloud solutions which could be effectively used for various Business purpose.