Explainable Artificial Intelligence for Predictive Maintenance Applications

Stephan Matzka
School of Engineering - Technology and Life
Hochschule für Technik und Wirtschaft Berlin
12459 Berlin, Germany
stephan.matzka@htw-berlin.de

***Abstract*—This paper presents and provides a realistic, yet synthetic, predictive maintenance dataset for use in this paper and by the community. An explainable model and an explanatory interface are described, trained using the dataset, and their explanatory performance evaluated and compared.**

Keywords—explainable artificial intelligence, predictive maintenance, dataset, decision tree, model agnostic method.

I. Introduction

Explainable Artificial Intelligence (XAI) is an ongoing effort in the field of artificial intelligence to augment machine learning methods and results with an explanation understandable by humans. The rationale behind this is, that in order for humans to trust a model's result, an explanation should be provided, as is usually the case in human interaction. Two main approaches to attain explanations understandable by humans are explainable models and explanatory interfaces (cf. [1]).

In recent years, intensive research has been conducted about the nature of a good explanation and about ways to generate these explanations from what are more often than not essentially black-boxes. An impressive overview about the topic's diverse aspects and methods is given in [2].

Current evaluation of trained models heavily relies on diagrams such as confusion matrices and receiver operating characteristic (ROC), or parameters such as precision, recall or F1-Scores, these are all statistical properties. These allow to show that a model, given data that is comparable to the used training-, validation- and test-sets, will behave with a predictable statistical performance on both known and previously unknown data.

What cannot be predicted by these evaluation methods is the performance of a trained model on a single data point. It is generally unclear what combination of factors is responsible for a given result and thereby unclear, if or when this result can be trusted (or not).

Of course, some machine learning methods, such as decision trees, are very easy to be interpreted my humans if the number of nodes is small and the features used at these nodes are known and understood. Note that using dimensionality reduction such as PCA or LDA (cf. [3]) on features violates the second requirement.

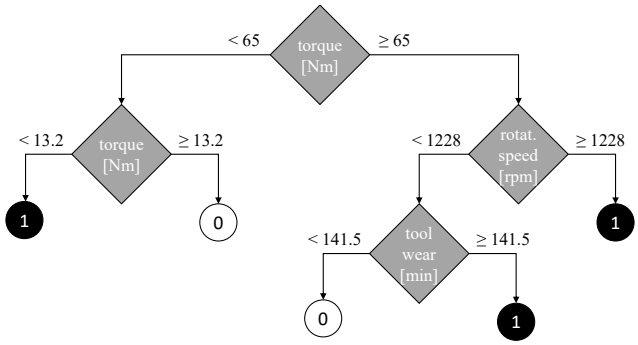

Fig. 1. Decision tree (#1) trained on the presented dataset using torque (in Nm), rotational speed (in rpm) and tool wear (in min) as nodes. A result of 0 represents normal operation, whereas a result of 1 represents a machine failure.

In Fig. 1 such a decision tree is shown, consisting only of four nodes with easily interpretable values ranging from the torque in Nm to the time the current tool has been in use (tool wear) in minutes. For trained maintenance personnel, this decision tree can be easily understood as: "small torque values under 13,2 Nm predict machine failure, as do high torque values over 65 Nm when combined with rotational speeds exceeding 1228 rpm or a tool wear longer than 141.5 minutes." As there is an immediate need to implement predictive maintenance methods in industry, e.g. [4], and the lack of trust in a model's accuracy is a major obstacle, explainable artificial intelligence may prove as the key factor to establish this technology.

In this paper we present and provide a synthetic, yet realistic, predictive maintenance dataset for use in our evaluation and by the community, seeing that predictive maintenance datasets are rarely made publicly available. Second, we describe, train and an explainable model and an explanatory interface and evaluate and compare the explanatory performance of both.

This paper is organized as follows: In section II we describe our dataset, in section III we train our classifiers and measure their performance. Section IV evaluates the explanations provided by an explainable model and an explanatory interface. Section V concludes this paper and an outlook on future work in this field is given.

978-1-7281-8701-3/20/$31.00 ©2020 IEEE
DOI 10.1109/AI4I49448.2020.00023

II. Predictive Maintenance Dataset

Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, we present and provide a synthetic dataset [5] that reflects real predictive maintenance data encountered in industry to the best of our knowledge and experience. The dataset consists of 10,000 data points stored as rows with 6 features in columns

1. product ID

   consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number.

2. air temperature [K]

   generated using a random walk process later normalized to a standard deviation of 2 K around 300 K

3. process temperature [K]

   generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.

4. rotational speed [rpm]

   calculated from a power of 2860 W, overlaid with a normally distributed noise

5. torque [Nm]

   torque values are normally distributed around 40 Nm with a $\sigma = 10$ Nm and no negative values.

6. tool wear [min]

   The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.

and a "machine failure" label that indicates, whether the machine has failed in this particular datapoint. The machine failure consists of five independent failure modes

1. tool wear failure (TWF)

   the tool will be replaced or fail at a randomly selected tool wear time between 200 – 240 mins (120 times in our dataset). At this point in time, the tool is replaced 74 times, and fails 46 times (randomly assigned).

2. heat dissipation failure (HDF)

   heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tool's rotational speed is below 1380 rpm. This is the case for 115 data points.

3. power failure (PWF)

   the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.

4. overstrain failure (OSF)

   if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 for M, 13,000 for H), the process fails due to overstrain. This is true for 98 datapoints.

5. random failures (RNF)

   each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for 19 datapoints, more frequent than could be expected for 10,000 datapoints in our dataset.

If at least one of the above failure modes is true, the process fails and the machine failure label is set to 1, which is the case for 339 datapoints. It is therefore not transparent to the machine learning method, which of the failure modes has caused the process to fail.

III. Classifier Training and Performance

While training an ideally fitted classifier for the dataset described in section II is not a focus of this paper, it never the less has to be done in order to have a complex classification model that can be understood by humans using the two explanatory models.

Our dataset, as inherently most predictive maintenance datasets, is severely imbalanced having only 339 datapoints labeled as machine failure. At the same time, a failure rate of 3.39 % would usually be considered a problematic process in mass production environments. The misclassification cost of a false negative is thus set to 30 times the cost of a false positive.

A. Complex Classifier Training and Performance

After initial evaluation and optimization of support vector machines, artificial neural networks we settle for a bagged trees ensemble classifier. This is to a certain extend intuitive as the database's rules for machine failure are a combination of thresholds in at least two features. The classifier's performance is shown in Table I and can be considered satisfactory for our purpose.

Another interesting result of this training is the estimated predictor importance (using the estimation metric described and implemented in [6]) of the provided features, as shown in Fig. 2, where it can be seen that torque and rotational speed dominate the classification process, whereas the temperatures and especially the type is not considered to be of much importance.

TABLE I. Confusion Matrix of the Bagged Trees Ensemble Classifier using 5-Fold Cross Validation.

		true class	
		failure	operation
predicted class	failure	294 (86.7 %)	45 (13.3 %)
	operation	121 (1.3 %)	9,540 (98.7 %)

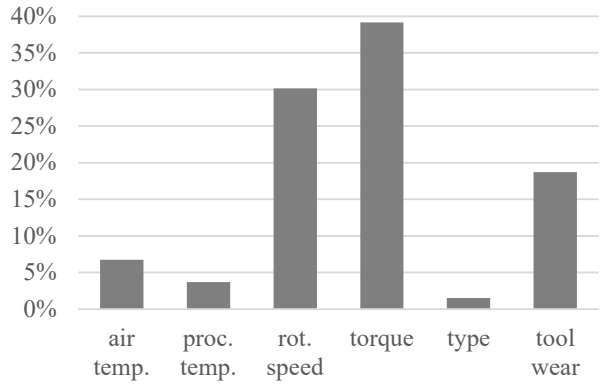

Fig. 2. Estimated relative predictor importance of the features used in the complex decision tree as described and implemented in [6].

B. Explainable Model Training

As an explainable model we train a set of 15 decision trees limited to a maximum of only 4 nodes for easy interpretability by a human. Each decision tree is trained using only 4 of 6 available features in the pattern shown in Table II. An example decision tree (number 1) is shown in Fig. 1.

TABLE II. FEATURES USED FOR THE TRAINING OF THE EXPLAINABLE DECISION TREES. BLANK FIELD INDICATE FEATURE WAS NOT PROVIDED, WHEREAS 0 INDICATES THAT PROVIDED FEATURE WAS NOT USED. A 1 INDICATES THAT THE FEATURE IS USED IN THE DECISION TREE.

tree num.	air temp.	proc. temp.	rot. speed	torq.	type	tool wear
1			1	1	0	1
2		0		1	0	1
3		0	1		1	1
4		0	1	1		0
5		0	1	1	0	
6	0			1	0	1
7	1		1		0	1
8	0		1	1		0
9	1		1	1	0	
10	0	0			0	0
11	0	0		1		1
12	0	0		1	1	
13	1	1	1			0
14	1	1	1		0	
15	1	1	1	1		
∑	5	3	10	10	2	6

It can also be seen in Table II that some features are used whenever they are available, i.e. rotational speed and torque and others only in few decision trees, such as type. To quantify this, we have summed up the used features in the last row of Table II. This closely mirrors the estimated importance of features for the bag of trees ensemble classifier shown in Fig. 2.

The different machine failure modes described in section II require individual feature combinations to be classified correctly, these are:

1. tool wear failure (TWF): tool wear
2. heat dissipation failure (HDF): air temperature, process temperature, and rotational speed
3. power failure (PWF): torque, and rotational speed
4. overstrain failure (OSF): torque, type and tool wear

therefore, some trees are inherently better equipped to correctly classify some failure mode than others. The percentage of used features for the respective failure mode is given in Table III.

TABLE III. PERCENTAGE OF USED FEATURES (CF. TABLE II) IN RESPECT TO REQUIRED FEATURES FOR A SPECIFIC FAILURE MODE.

tree num.	TWF	HDF	PWF	OSF
1	**100%**	33%	**100%**	**67%**
2	**100%**	0%	50%	**67%**
3	**100%**	33%	50%	**67%**
4	0%	33%	**100%**	33%
5	0%	33%	**100%**	33%
6	**100%**	0%	50%	**67%**
7	**100%**	**67%**	50%	33%
8	0%	33%	**100%**	33%
9	0%	**67%**	**100%**	33%
10	0%	0%	0%	0%
11	**100%**	0%	50%	**67%**
12	0%	0%	50%	**67%**
13	0%	**100%**	50%	0%
14	0%	**100%**	50%	0%
15	0%	**100%**	**100%**	33%

It can be seen in Table III, that for all failure modes besides overstrain failure (OSF) there exist feature combinations that cover 100% of the required features to detect the failure. For OSF, at most two of three features were used, which can be attributed to the low estimated predictor importance of the *type* feature (cf. Fig. 2), resulting in type used only in two explainable decision trees (cf. Table II).

IV. EVALUATION OF EXPLAINATORY PERFORMANCE

In this paper, two different approaches are used to explain the ensemble classifier's results. First, we use the explainable decision trees presented in section III-B, and second, we use the normalized feature deviation as explanatory interface. For this we use a subset of 20 random datapoints given in Table IV, stratified to include five instances of each failure mode, at least two L *types*, and one M and H *type* for each failure mode, and altogether 50% L *types*, 30% M *types*, and 20% H *types*.

TABLE IV. DATAPOINTS USED FOR EVALUATION OF EXPLANATORY PERFORMACE FOR DIFFERENT FAILURE MODES (TWF TO OSF). PTC DENOTES THE BAG OF TREES CLASSIFICATION RESULT.

UID	prodID	TWF	HDF	PWF	OSF	BTC
2672	M17531	1				0
3866	H33279	1				1
6341	H35754	1				0
8358	L55537	1				0
9019	L56198	1				0
3237	M18096		1			0
4079	H33492		1			1
4174	M19033		1			1
4327	L51506		1			1
4502	L51681		1			1
464	L47643			1		1
1493	M16352			1		1
3001	H32414			1		1
7537	L54716			1		1
8583	M23442			1		1
250	L47429					1
3020	L50199				1	1
5400	H34813				1	1
7592	M22451				1	1
9660	L56839				1	1

It can be seen in Table IV in the BTC column, that the Bag of Trees classifier described in section II-A has difficulties with the tool wear failure detection. This can in part be attributed to the low number of TWF cases in the dataset and the random nature of timely tool replacement or tool failure, a frequent observation (and obstacle for machine learning) in real predictive maintenance situations.

A. Explainable Model Performance

We provide the datapoints given in Table IV to the 15 explainable decision trees described in section III-B. If more than one decision tree returns a positive result, then the tree using the features with the highest sum of estimated predictor importance (cf. Fig. 2) is chosen as an explanation. The user output then is the path leading towards the positive result for the given feature values.

In Table V we see that the bagged trees classifier's limitation in detecting the *tool wear* failure mode reflects on the explainable decision trees, as not a single decision tree classifies any of the cases as failure. Examining these datapoints, this is due to the average torque values, a feature that is used in almost all decision trees. However, tool wear failures with comparably high or low torque values also qualify for the power failure and overstrain failure mode and have thus been disregarded for the testcases in Table IV.

Also, three HDF failures out of five are not detected using the explainable decision trees. This is a direct result of using decision trees with a low number of nodes to ensure an explanatory value. However, the difference between air- and process temperature is, besides rotational speed, the key factor here. As no temperature difference feature is provided, this case is inherently difficult for this class of classifiers.

After these sobering findings, we can report that that for all other datapoints (besides UID 1493, an edge case) at least partially useful (2 cases), but mostly very (9 cases) useful explanations were provided by the explanatory decision trees.

B. Explanatory Interface

As a second method, we use the normalized feature deviation as a simple model independent explanatory interface. For this, each feature is normalized to an expected value of 0 and a standard deviation of 1. Table VI shows this for the first instance of each failure mode, with the absolute feature values in the first row and the feature value after normalization in the second row of each datapoint.

The user is then provided with an explanation based upon the feature, or a small number of features, with the maximum absolute deviations. For UID 2672 in Table VI and two features

TABLE V. SCORED EXPLAINATION OF DECISION TREES FOR SELECTED DATAPOINTS (CF. TABLE IV). SCORES (SCR) RANGE FROM ++ (VERY), + (PARTIALLY), − (LIMITED), TO − − (NO) USEFUL EXPLAINATION. THE SELECTED TREE IS INDICATED BY BOLD TYPE.

UID	Mode	BTC	Trees	Explanation	Scr
2672	TWF	0		none	n/a
3866	TWF	1		none	− −
6341	TWF	0		none	n/a
8358	TWF	0		none	n/a
9019	TWF	0		none	n/a
3237	HDF	0		none	n/a
4079	HDF	1		none	− −
4174	HDF	1		none	− −
4327	HDF	1	13-**15**	torque < 65 Nm rotSpeed < 1380 rpm airTemp < 301.5 K procTemp < 310.5 K	+ +
4502	HDF	1			+ +
464	PWF	1	1-**8**,9, 11-15	torque < 13.2 Nm	+
1493	PWF	1		none	− −
3001	PWF	1	1,2,4, 5,6,**8**, 9,11, 12,15	torque > 65 Nm rotSpeed > 1229 rpm	+ +
7537	PWF	1	1-5, 6-**8**, 9-15	torque < 13.2 Nm	+
8583	PWF	1	1,2, 4-6,**8**, 9,11, 12,15	torque > 65 Nm rotSpeed > 1229 rpm	+ +
250	OSF	1	2,6, **11**	torque > 53.6 Nm tool wear > 194.5 min	+ +
3020	OSF	1	2,3, 6,**11**		+ +
5400	OSF	1	2,6, **11**		+ +
7592	OSF	1	2,6, **11**		+ +
9660	OSF	1	2,3, 6,**11**		+ +

TABLE VI. ABSOLUTE FEATURE VALUES (FIRST LINE) AND NORMALIZED FEATURE DEVIATIONS IN SIGMA (SECOND LINE). IN EACH, THE TWO MAXIMUM ABSOLUTE DEVIATIONS ARE MARKED BOLD.

UID	Mode	air temp.	proc. temp.	rot. speed	torq.	tool wear
2672	TWF	299,7	309,3	1399	41,9	221
		-0,15	-0,47	**-0,78**	0,19	**1,78**
3237	HDF	300,8	309,4	1342	62,4	113
		0,40	-0,41	**-1,10**	**2,25**	0,08
464	PWF	297,4	308,7	2874	4,2	118
		-1,30	-0,88	**7,45**	**-3,59**	0,16
250	OSF	298	308,3	1405	56,2	218
		-1,00	-1,15	-0,75	**1,63**	**1,73**

this would be: "The classification as *fail* is due to a high tool wear of 221 mins and a low rotational speed of 1399 rpm". In this case, the first information is essential whereas the second information is irrelevant (but not misleading). The score for this would therefore be "partially useful explanation" (+).

The explanations in Table VII are mostly partially useful (14 cases), as often only one of two explanations is relevant. Very good explanations have been provided in 6 cases. Inherently there are no missing explanation as opposed to the explanatory decision trees.

C. Quantitative Comparison of Scored Explanations

A quantitative comparison of the explainable decision trees and the normalized feature deviations is given in Fig. 3.

Fig. 3 shows, that using normalized feature deviations as explanatory interface results in mostly partially useful explanations, fewer very good results and only to limited useful explanations in our dataset. Using decision trees as explainable models in contrast yields mostly very useful explanations, but also provides no useful explanation in four cases and another five cases are not considered (n/a) due to the lack of a correct classification by the complex ensemble classifier.

A possible solution could be to present the user with the decision tree, if an explanation can be extracted, as they tend to have a higher quality than normalized feature deviations. In case no explanation is provided by the decision trees, normalized feature deviations proved to provide partially or very useful explanations for all testcases, and thus present a consistently good alternative.

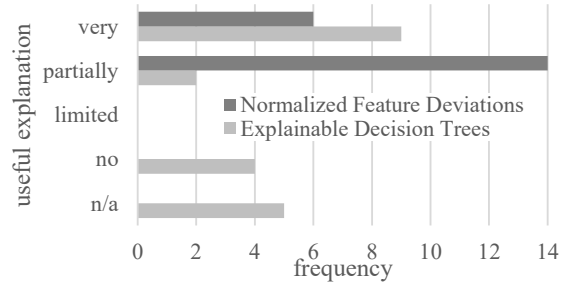

Fig. 3. Quantitative comparison of explanation scores between explainable decision trees and normalized feature deviations.

V. CONCLUSION AND FUTURE WORK

In this paper, two methods to provide an explanation for the classification result of a complex ensemble classifier are evaluated on a synthetic predictive maintenance dataset. Both methods have inherent strengths and weaknesses, but provide an overall benefit for the user without high additional costs.

The explanations provided by the decision trees tends to be of a higher quality, but in a considerable number of cases do not provide any explanation. On the other side, normalized feature deviations provide explanations of a consistent, yet slightly lower, explanatory quality.

TABLE VII. SCORED EXPLANATION OF NORMALIZED FEATURE DEVIATION FOR SELECTED DATAPOINTS (CF. TABLE IV). THE SCORES (SCR) ARE DESCRIBED IN TABLE V.

UID	Mode	BTC	Explanation	Scr
2672	TWF	0	high tool wear of 221 mins low rot. speed of 1399 rpm	+
3866	TWF	1	high tool wear of 228 mins high air temp. of 302.6 K	+
6341	TWF	0	high tool wear of 210 mins low rot. speed of 1397 rpm	+
8358	TWF	0	high tool wear of 210 mins low rot. speed of 1397 rpm	+
9019	TWF	0	high tool wear of 217 mins low air temp. of 297.3 K	+
3237	HDF	0	high torque of 62.8 Nm low rot. speed of 1342 rpm	+
4079	HDF	1	high torque of 62.8 Nm low rot. speed of 1294 rpm	+
4174	HDF	1	high air temp. of 302.2 K low rot. speed of 1346 rpm	+ +
4327	HDF	1	high torque of 55.8 Nm low rot. speed of 1362 rpm	+
4502	HDF	1	high torque of 54.0 Nm low rot. speed of 1307 rpm	+ +
464	PWF	1	high rot. speed of 2874 rpm low torque of 4.2 Nm	+
1493	PWF	1	high torque of 58.5 Nm low air temp. of 298 K	+
3001	PWF	1	high torque of 72.8 Nm low rot. speed of 1324 rpm	+
7537	PWF	1	high rot. speed of 2579 rpm low torque of 12.5 Nm	+
8583	PWF	1	high torque of 72.8 Nm low proc. temp. of 308.1 K	+
250	OSF	1	high tool wear of 218 mins high torque of 56.2 Nm	+ +
3020	OSF	1	high tool wear of 207 mins high torque of 54.2 Nm	+ +
5400	OSF	1	high tool wear of 218 mins high air temp. of 302.8 K	+
7592	OSF	1	high torque of 61.3 Nm high tool wear of 202 mins	+ +
9660	OSF	1	high torque of 61.9 Nm high tool wear of 216 mins	+ +

Providing both explanations is discouraged in literature, as the possible increase in information is far outweighed by the confusion caused by inconsistent explanations [2].

Future work will include extending this evaluation using local interpretable model-agnostic explanations (LIME [7]), and developing a method to limit the "no useful explanation" result of the explainable decision trees, e.g. by gradually lowering the threshold for a true positive at the expense of an increased number of false positives.

REFERENCES

[1] D. Gunning, "Explainable artificial intelligence (xai)." *Defense Advanced Research Projects Agency (DARPA),* 2017

[2] C. Molnar, "Interpretable machine Learning", https://christophm.github.io/interpretable-ml-book/ , 2020.

[3] G. McLachlan, "Discriminant analysis and statistical pattern recognition". Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 2004.

[4] S. Matzka, "Using Process Quality Prediction to increase Resource Efficiency in Manufacturing Processes", First International Conference on Artificial Intelligence for Industries (AI4I 2018), 2018

[5] S. Matzka, "AI4I 2020 Predictive Maintenance Dataset", www.explorate.ai/dataset/predictiveMaintenanceDataset.csv, submitted to UCI Machine Learning Repository, 2020.

[6] Mathworks, "Estimates of predictor importance for classification ensemble of decision trees" https://de.mathworks.com/help/stats/compactclassificationensemble.predictorimportance.html, 2020.

[7] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.