Performance of Explainable AI Methods in Asset Failure Prediction

Jakub Jakubowski[1][0000−0002−4773−9086], Przemysław Stanisz, Szymon Bobek[2][0000−0002−6350−8405], and Grzegorz J. Nalepa[2][0000−0002−8182−4225]

[1] AGH University of Science and Technology, 30-059 Krakow, Poland
[2] Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI), Institute of Applied Computer Science, Jagiellonian University, 30-348, Krakow, Poland

Abstract. Extensive research on machine learning models, which in the majority are black-boxes, created a great need for the development of Explainable Artificial Intelligence (XAI) methods. Complex machine learning (ML) models usually require an external explanation method to understand their decisions. The interpretation of the model predictions are crucial in many fields, i.e., predictive maintenance, where it is not only required to evaluate the state of an asset, but also to determine the root causes of the potential failure. In this work, we present a comparison of state-of-the-art ML models and XAI methods, which we used for the prediction of the RUL of aircraft turbofan engines. We trained five different models on the C-MAPSS dataset and used SHAP and LIME to assign numerical importance to the features. We have compared the results of explanations using stability and consistency metrics and evaluated the explanations qualitatively by visual inspection. The obtained results indicate that SHAP method outperforms other methods in the fidelity of explanations. We observe that there exist substantial differences in the explanations depending on the selection of a model and XAI method, thus we find a need for further research in XAI field.

Keywords: machine learning · explainable artificial intelligence · predictive maintenance.

Acknowledgements This paper is funded from the XPM (Explainable Predictive Maintenance) project funded by the National Science Center, Poland under CHIST-ERA programme Grant Agreement No. 857925 (NCN UMO-2020/02/Y/ST6/00070).

1 Introduction

In the last ten years, we observe a tremendous growth of products and research papers related to machine learning and data mining. The increasing number of real-life applications of these techniques is driven by multiple factors. From the technical perspective, cloud computing allows to train complex models on

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

specially designed clusters. Progress in the field of big data enables researchers to store and process an enormous amount of data. Development of machine learning techniques, especially deep learning, affected in the improvement of the model accuracy with the significant growth in areas like image classification, natural language processing, speech recognition, decision making, and time series analysis. Furthermore, artificial intelligence (AI) models are now much more accessible thanks to the development of open source frameworks in a variety of programming languages including Python, R, and C++.

For manufacturing companies, artificial intelligence is one of the driving factors in the transition to Industry 4.0 [4], which is the key challenge in the years to come. The applications of artificial intelligence models in the manufacturing industry include process control, production planning, network traffic monitoring, and predictive maintenance (PdM). The potential benefits for companies to use machine learning for monitoring the health of their assets are very high as nowadays relatively simple techniques are used. Most of the machine's equipment is now replaced either in a corrective or preventive manner. The first aims to replace the element after its failure, whereas the goal of the second method is to replace an element after a predefined period of time, before it fails. Both strategies have significant drawbacks, which justifies the need for new solutions. The corrective approach is not suitable for critical assets, which failure may cause safety issues or significant financial losses. On the other hand, the preventive approach may increase the total operating cost due to the increased frequency of asset replacements. Using machine learning for estimating the condition of the machine may be a promising alternative, which can give the industry substantial gains. The topic of RUL prediction with the use of artificial intelligence approaches was widely studied by researchers in the last years, especially in the field of deep learning [23].

One of the major issues with machine learning models is their black-box nature, which impedes the understanding of the model and the result. This lack of transparency may impact the trustworthiness of the model during the development phase as well as during its operation in the production environment (real-life applications), especially when there is a need to understand factors influencing the model decision. Except for that, there are also legal concerns, which may oblige companies and institutions to provide explanations for the model prediction whenever it affects user [6].

To address these issues, Explainable AI (XAI) methods, which try to explain the prediction of black-box models, gained popularity among researchers in recent years. Despite the rapid growth in the field of XAI, there are concerns about its efficiency in giving the right explanations, which lead to the conclusion that black-box models should not be used in any high-stake decisions [20]. Another alternative is to use glass-box models, which are models inherently interpretable, thus they do not require any additional mechanism to provide the explanations.

In this research, we evaluate the performance of black-box explainablity methods and compare it with the results obtained by the interpretable machine learning model. We focus on two explainability metrics - stability and consis-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

tency. Those metrics can be used for quantitative assessment of the explanations produced by different models. They allow to compare the explanations within one model and between different models. This paper is a part of Explainable Predictive Maintenance (XPM) project, which is devoted to the use of XAI in predictive maintenance solutions. In the project we focus on four different real-life cases, which namely are: steel manufacturing, city subway, wind farms and trucks maintenance. This paper constitutes a preliminary work in the area of evaluation of XAI methods for PdM. Hence, to assure reproducibility, we have based this study on a public data set, which describes a degradation of turbofan engine (CMAPSS) (provided by NASA). In particular, we aim at analysing which XAI methods and ML models are suitable to predict the RUL of the tubrofan engine and provide acceptable explanations of that decision.

The rest of the paper is organized as follows. In Section 2, we provide an overview of state-of-the-art machine learning, explainability methods, interpretable models, and metrics used for the evaluation of model explanations. In Section 3, we present the failure prediction case, which we use in the study – this is the dataset coming from the simulations of an aircraft turbofan engine. We also present our approach towards predicting asset failure with explanations and evaluate the quality of those explanations. In Section 4, we present the results obtained in this study and in Section 5 we summarize our research and point out directions for further investigation.

2 Explainable Artificial Intelligence

Machine learning models combine complex mathematical algorithms with the data coming from a certain process to build a general mathematical model, which is able to make a correct prediction on previously unseen data. In most cases, complex models are able to achieve very high accuracy scores in the certain problem, but generally they are significantly more difficult to explain [7].

2.1 Explaining black-box models

Explainable AI algorithms are able to build an understanding of the black-box models by applying different methods, i.e., input perturbations, to find the driving factors of the prediction. The process of producing explanations depends on factors like the characteristics of the explained model, data structure, and prediction type, i.e., image data need different methods of explanation than tabular data as the values of each pixel are not understandable by humans straightaway – they need to be visualized.

The methods might be either model-specific or model-agnostic. In the first case, only a specific predefined type of models can be explained – examples of such methods are Grad-CAM [22], which is designed to give visual explanations of deep learning models, and RFEX, which focuses strictly on the explanation of Random Forest Classifier [17]. The model-agnostic approach is not based on selected AI methods, but aims to build framework for explanations of any model.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

Examples of such explanation methods are Local Interpretable Model-Agnostic Explanations (LIME) [18], SHapley Additive exPlanations (SHAP) [12], Anchors [19].

Another way of dividing explainability methods is into local and global explanations. Local explanations aim to understand why the model made a certain prediction in a selected case (one observation), while the global explanations try to give an overall understanding of the model as a whole.

The XAI methods may also differ based on the way of presenting the explanation to the end-user. SHAP gives a numerical value of feature importance, which tries to evaluate how the prediction of the model changes under the certain condition (value of the selected feature), while Anchors explain the prediction in the form of rules.

2.2 Glass-box models

The problem of explainability does not exist in the case of glass-box models, which are interpretable without further need of using explainability methods [20]. This may increase the reliability of the machine learning model, because the explanations do not rely on the external method (such as SHAP), which may also be not trustworthy for the end-user of the solution.

One of the simplest and most widely used interpretable models are linear models (i.e., linear regression, logistic regression). However, their performance is known to be relatively poor on more complex data sets. An extension of the linear models, which may increase their accuracy are, for example, Generalized Additive Models (GAMs) [8], which instead of using linear relationships between features, use many nonlinear equations, which are summed to give the prediction. Comparing them with linear models trade-off between accuracy and explainability is observed – at the cost of higher performance, GAMs are less interpretable. Moreover, their accuracy is generally not as high as state-of-the-art algorithms [13].

A promising algorithm, which tries to achieve high accuracy and interpretability is Explainable Boosting Machine (EBM) [15]. It is based on the idea of General Additive Models, but uses more advanced machine learning techniques like boosting and bagging to improve the accuracy of the model.

2.3 Explainability metrics

In the previous sections, we have highlighted that even though explainability methods give more insight into the prediction of the model, they may not be trustworthy themselves. Thus, there is a need to derive metrics, which can be used to validate the performance of the models and XAI methods. The two base requirements, which must be fulfilled to perceive model explanations as trustworthy are that (1) similar observations should lead to similar explanations within a certain model and (2) the explanations for a given observation should be similar irrespective of the machine learning model and explanation method.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

The first criterion assures that small changes in the observations will not lead to high changes in the explanations. This is referred as stability or robustness of the explainable model. Alvarez-Melis and Jaakkola have proposed to use a metric based on Lipschitz continuity to calculate the stability of the explanation at a given point [2]:

$$\tilde{L}_x(x_i) = \max_{x_j \in \mathcal{N}_\in(x_i) \leq \epsilon} \frac{\|\Phi_i - \Phi_j\|_2}{\|x_i - x_j\|_2} \tag{1}$$

where $\tilde{L}_x(x_i)$ is the stability of the point x_i, Φ_i and Φ_j are the feature importance vectors of points x_i and x_j (each feature of an observation has a feature importance assigned to it by XAI method), $\mathcal{N}_\in(x_i)$ is the neighborhood of x_i, which is defined as all points which have distance (defined as L2 norm) to x_i smaller than ϵ.

The general idea behind this metric is to find all points in the neighborhood of the point x_i and find the maximum dissimilarity, which is defined as the Euclidean distance between explanations divided by Euclidean distance between the points. The lower the value of stability, the better performance of the model at a given point. The major drawback of this metric is that its value is relative, therefore it is not possible to conclude on the stability of one model without having a comparison with other models.

The second criterion is used for the comparison of different models and validate their consistency with each other. It assumes that if we have two different models (or explanation methods) then we expect to have similar explanations for the same observation. If this is not the case, then either one model (at least) is not making a good prediction or the explanation method is not properly finding relevant features. The consistency metric has been proposed in [3]. For the comparison of two different explanations, equation takes the following form:

$$C(\Phi_{1,i}, \Phi_{2,i}) = \frac{1}{\|\Phi_{1,i} - \Phi_{2,i}\|_2 + 1} \tag{2}$$

where $C(\Phi_{1,i}, \Phi_{2,i})$ if the consistency of i^{th} observation, $\Phi_{1,i}$ and $\Phi_{2,i}$ are the feature importance vectors of i^{th} observation for the first and second model respectively.

In a perfect scenario, when all feature importances are equal, the consistency is equal to 1 and it drops as the distance between explanations increases. Theoretically, the lowest possible value of consistency defined in such a way is 0. Nevertheless, the value of consistency is also dependent on the magnitude of the feature importance vector, therefore it may be affected by feature engineering, i.e., scaling.

3 Asset failure prediction

3.1 C-MAPSS data set

Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) [5] is a software developed by National Aeronautics and Space Administration (NASA)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

to simulate the behaviour of turbofan engines. Based on this tool, Saxena et al. [21] have prepared a dataset, which consists of run-to-failure simulations of hundreds of turbofan engine units.

The dataset consists of 21 features, i.e., temperature, pressure, fan speed, fuel and coolant flow, and 3 operational parameters (settings). Each observation is the average measurement from a simulated flight of a single unit. As the number of completed cycles (flights) of a certain unit increases, the gradual degradation is observed. The turbofan engines may operate with different external conditions (up to six) and exhibit one of two failures – high-pressure compressor (HPC) or fan degradation. The simulation dataset is divided into four subsets named FD001-FD004, which contain data of different complexity – from simpler (one external condition and one type of failure) to more complex (six external conditions and two types of failure) cases.

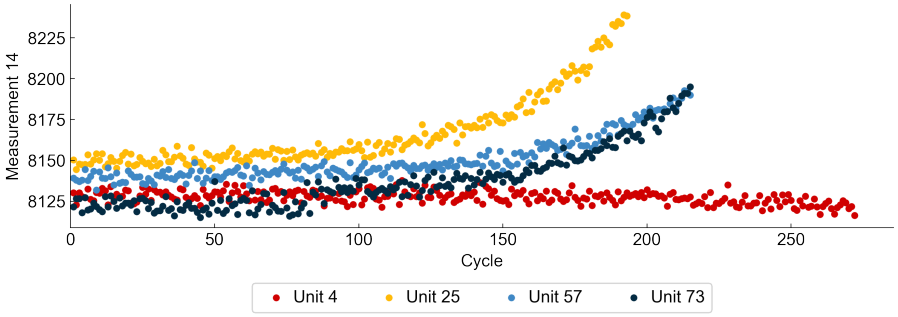

Fig. 1. Exemplary plot showing how the value of selected feature changes with the progress of engine degradation (FD003).

For each turbofan engine, the RUL at a given point may be determined based on the total number of completed cycles. This way, a prediction model, which determines the state of health of the unit may be developed. Figure 1 presents how the selected measurement deviates from normal working conditions as the number of cycles increases and the turbofan engine deteriorates. Figure 2 shows how the distributions of some measurements differ in normal working conditions and at the end of life. In most cases, a shift towards lower or higher values is observed as the unit undergoes failure.

3.2 Failure prediction

The prediction of unit failure may be considered as a problem of finding the remaining useful life (RUL) estimation of the unit given the current working conditions. The RUL prediction on the C-MAPSS dataset has been widely studied and multiple prediction models were proposed. Most recent research is mainly

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

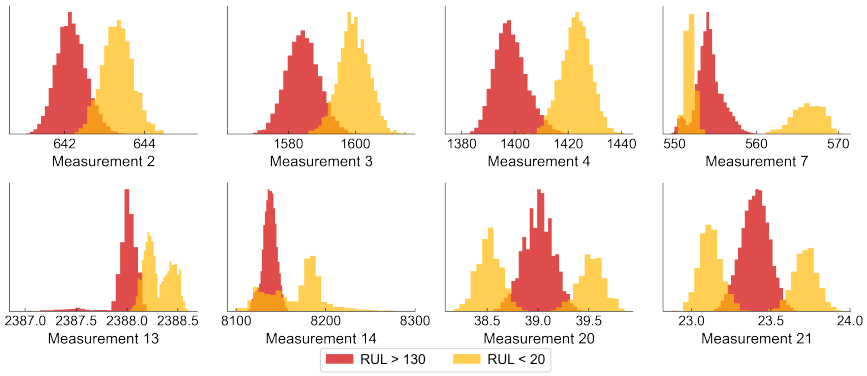

Fig. 2. Comparison of selected measurements distribution in normal (RUL $>$ 130) and degradation (RUL $<$ 20) conditions (FD003).

focused on the development of different deep learning architectures [10, 16, 11, 1], but more classical ML approaches were also studied [14, 9].

Our work is dedicated mostly to the topic of Explainable AI, rather than the development of new model architectures, thus we have used more commonly used state-of-the-art techniques, which are known for their performance on various types of data and prediction problems. The algorithms selected for the prediction task are XGBoost (XGB), Random Forest (RF), Support Vector Machine (SVM), and Multi-layer Perceptron (MLP). We also use Explainable Boosting Machine [15] as a glass-box alternative for the models listed above.

The dataset on which we have trained the models is the FD003, which contains a simulation of 100 turbofan engine units in a run-to-failure manner with over 24,000 observations in total. Each simulated flight is made under the same external conditions with two possible types of failure. For each unit, we calculate the RUL based on the known value of the cycle for each observation and known total number of cycles for a given unit. The feature scaling is performed in the following manner – for each unit we take the first 50 cycles (which are assumed to be always healthy working conditions) and scale all features linearly to the [0, 1] range. Then we apply the obtained scale to all observations of the given unit. With such an approach, we are able to find the relative change of the measurements in relation to the baseline, which are the first 50 cycles. The first 50 cycles are then removed from the training and test data sets, as in practice when we calculate RUL for a new unit, those points could not be predicted, as the baseline is not yet known. We also apply a rectification of RUL, which is a common practice in the case of such problems. Whenever the value of RUL exceeds 130, it is limited to 130. This is required, because when RUL is higher than 130 cycles, there are no signs of asset degradation and thus no model is able to precisely distinguish between, i.e., RUL=300 and RUL=130.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

For every model, we conduct hyperparameter tuning using a grid search method to find a set of parameters, which assure the high accuracy of the model. The model accuracy is determined with the root mean square error ($RMSE$) as the model evaluation metric. The cross-validation technique with 5 folds is used to eliminate the possibility of model overfitting – data is divided into train and test sets in such a way that every unit can be only in one of these two sets at a time.

3.3 Failure explanation

After training the models with the best found hyperparameters, we use SHAP and LIME methods to get explanations of every model for several randomly selected units.

Then, we compare the explanations in several manners. As the evaluation of the robustness, we determine the stability range for every model and compare them together. To evaluate the stability in a more qualitative manner, we plot how the feature importance for different models changes as RUL decreases in each selected unit. We expect that for the explanations to be trustworthy, the explanation for the RUL prediction should be similar throughout the whole degradation process. Otherwise, the end-user of the XAI model may be misguided and lose trust in the predictions.

We also calculate the mean consistency between every combination of the two models to check which models give the most consistent explanations.

Additional issue in the comparison of the explanation methods is that the feature importance values produced by different XAI methods cannot be directly compared, as the meaning of the feature importance magnitude might be different. To overcome this problem, we have scaled the feature importances for each method in a following manner:

$$a_E = P_{95}(\|x_i\| : \{x_i \in X_E\}) \tag{3}$$

$$\Phi^s_{E,m,i} = \frac{\Phi_{E,m,i}}{a_E} \tag{4}$$

where a_E is the scaling factor for an explanation method E, $P_n(\mathrm{X})$ denotes the n^{th} percentile of multiset X, X_E is the concatenated multiset with all the feature importances of a given explanation method (for all models) and $\Phi_{E,m,i}$ is the feature importance of i^{th} observation for a ML model m and explanation method E – superscitpt s denotes a scaled value.

We have decided on the 95^{th} percentile to minimize the effect of the outliers on the final results. The scaling factors were determined for each method without distinguishing between the models to preserve the same scale within the explanation method, i.e., we assume SHAP values for all models are comparable without scaling. Although feature importance in all explanation methods may be positive or negative, in all cases $\Phi_i = 0$ means that a certain feature has no impact on the result. Thus, it is important to assure during scaling that this point does not shift, what could be achieved with, i.e., a simple min-max normalization.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

4 Results and discussion

In this section, we present the results of our study. We have trained five different models to predict the remaining useful life on the FD003 dataset. The models were trained and evaluated using Python programming language and sckit-learn library. Table 1 presents the root mean square error ($RMSE$) and coefficient of determination (R^2) calculated on the test dataset for all models used in this study.

Table 1. Metrics of machine learning models on FD003 dataset.

	XGB	**RF**	**SVM**	**MLP**	**EBM**
RMSE	**14.8**	15.0	15.1	15.1	15.4
$\mathbf{R^2}$	0.878	0.867	0.884	**0.887**	0.862

All models achieved comparable performance on the test dataset and their accuracy is acceptable to use them for the prediction problem defined. From the test dataset, we have randomly selected 10 units and produced explanations for them with the use of SHAP and LIME methods – in contrast to Explainable Boosting Machine algorithm, which is interpretable. We have only explained the observations, which had an actual RUL below 130, as we are particularly interested in the explanations for the low RUL values – the explanations for the healthy turbofan unit have no practical significance.

In Figure 3 we present how the feature importance values of a selected measurement changes as RUL decreases. SHAP values reflect the progressing degradation process, however for SVM, XGB, and RF the explanation scores converge to a certain value, while for MLP a decrease is still observed. Based on the behaviour of the measurement, MLP response seems to be more intuitive. In the case of LIME explanations, we observe three intervals: stability at high RUL values, fluctuations at intermediate RUL values, and stability at lower RUL values. Such a response does not seem to be of practical use, as in the intermediate RUL interval the state of the engine is not well explained. Changes in the explanation scores of EBM are similar to SHAP values, however higher fluctuations are observed.

The distribution of stability (as defined in equation 1 – with $\epsilon = 2.0$) is presented in the Figure 4. The best stability was obtained for the models explained with the SHAP method . The lowest median was achieved by Multi-layer perceptron model, nevertheless the results for other models (explained with SHAP) are comparable.

In the Figure 5 we presents the mean consistency between each model. The results are relatively far from 1.0, which may indicate the models are very far from being consistent, which raises the issue of their fidelity. The highest consistency score is observed between the three pairs of models: XGB and RF with

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

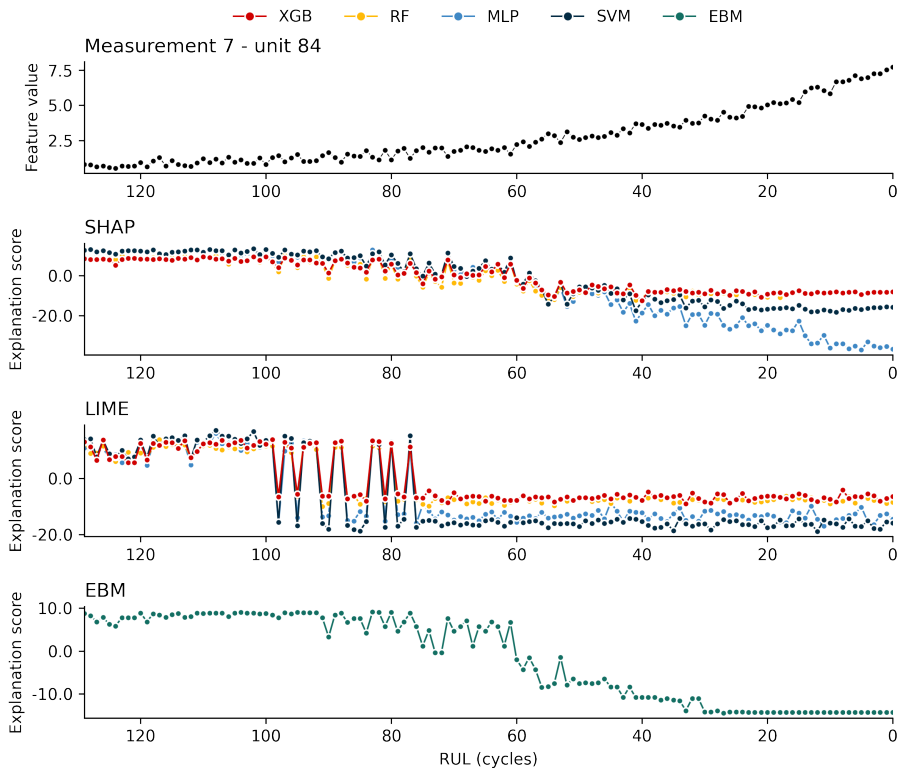

Fig. 3. Explanation scores of the Measurement 14 for a randomly selected unit.

SHAP, XGB and RF with LIME, SVM and MLP with SHAP. This leads to observations that consistency is higher if we compare two models using the same XAI method and that kind of model (tree-based or not) also impacts the consistency. The explanations of the same model with different XAI techniques give lower consistency than in the case of the pairs mentioned above. It shows that the selection of important features is not only driven by the model training, but is also dependent on the choice of the XAI method.

In the analyzed dataset, each unit undergoes one of the two failures, which should also be visible in measurements and explanations – different measurements may impact the RUL depending on the type of failure that is occuring in the engine. Thus, we expect to have two clusters of failure data, each having observations from one failure. Those clusters should be both visible in the measurements as well as explanations. In the Figure 6 we present the visualization of all features and explanations for the MLP model by reduction of the data dimensionality with the use of the Principal Component Analysis (PCA) method. The visualization of the measurements implies that we are dealing with

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

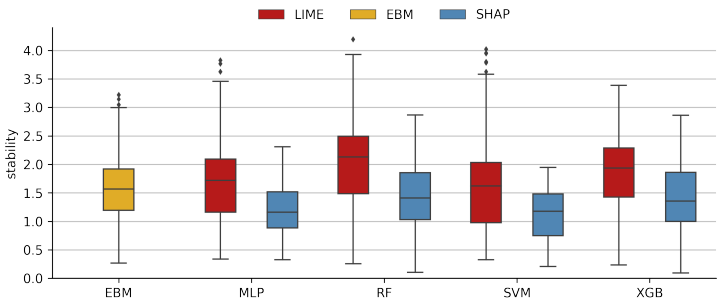

Fig. 4. Stability distribution of the investigated models for 10 randomly selected turbofan units. The lower is the median and the variance, the better.

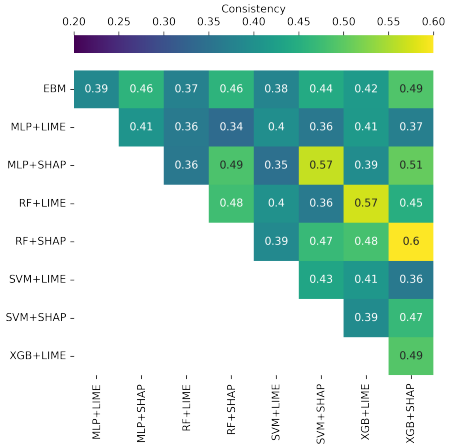

Fig. 5. Mean consistency between the models for 10 randomly selected turbofan units.

two distinct failures. There is a common starting point (normal condition) and as the RUL decreases, the measurements move in different directions of the plane. The behaviour of the SHAP values is similar to measurements, which shows that there is a consistency between them. On the other hand, in the case of LIME, four distinct regions are present, which cannot be simply explained by the distribution of the dataset. Nevertheless, it is still a noticeable shift between normal and failure points. This may be driven by the fact that LIME is known to be affected by small perturbations in the dataset [2].

5 Conclusion and future works

In this paper, we have discussed the problem of Explainable AI in the predictive maintenance case. We have focused on the comparison of feature importances

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

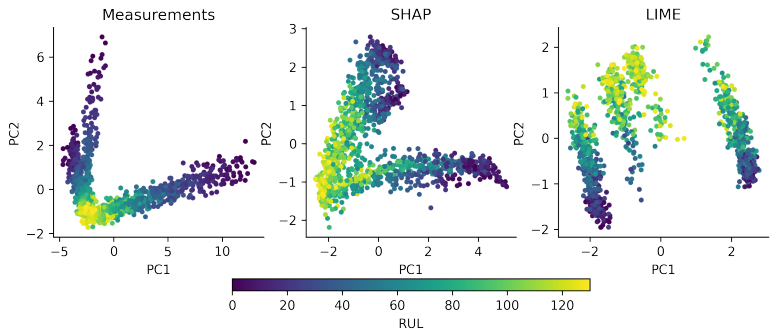

Fig. 6. PCA visualization of the measurements and explanations.

assigned by SHAP and LIME explainability methods for different types of black-box models. We also used Explainable Boosting Machine, which is a glass-box model and can provide explanations without the utilization of external explainability methods.

The results show that the SHAP method outperforms LIME and EBM in terms of stability and consistency of the explanations. The most promising results were obtained with MLP model, where the response of the XAI model was in our opinion the most reasonable, which was also confirmed by one of the best results in the stability metric. We have observed that the LIME method is not performing well in terms of stability and its results might be affected by small changes in the model, which is coherent with other studies. Explainable Boosting Machine has shown comparable performance in terms of prediction accuracy as the other techniques used (it was slightly less accurate than the rest of algorithms), but the stability of the explanations was worse than in the case of SHAP method. This implies that the glass-box models may not always perform better in terms of stability than a combination of black-box with XAI method.

The study has shown that there exist differences between the explanations that depend on the XAI method and ML algorithm. Not only different models result in different feature importances assigned by XAI methods, but there also exists a disagreement between XAI explanations for the same model. This indicates that the differences in the XAI methods are not coming only from the internal decisions of ML model, but also from the XAI methods themselves. Thus, there is a need for further research on the machine learning models and XAI methods, which will lead to the production of accurate and trustworthy algorithms for predictive maintenance tasks. The situation when unstable and inconsistent algorithms are used in the production environment may lead to the loss of trustworthiness of ML models by the end-users of those algorithms.

In future work, we plan to further investigate the topic of XAI methods in predictive maintenance applications with a special focus on remaining useful life estimation. Our next works will be devoted to the real-life use cases we plan to

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

investigate in the Explainable Predictive Maintenance Project. We want to focus on the explainability problem in more complex deep learning architectures, which show promising potential in PdM use cases, i.e., convolutional, LSTM, Transformer Networks or ensembles of them. We also see a great need for further research on explainability metrics – the current metrics give some valuable information. However, we observe that they cannot be evaluated in a straightforward way, i.e., they depend on the magnitude of the feature importance vectors, and they do not provide information on the source of bias in explanations. We also plan to investigate more in-depth differences between the explanations of SHAP, LIME, EBM and others explanation methods.

References

1. Abid, K., Sayed-Mouchaweh, M., Cornez, L.: Deep ensemble approach for rul estimation of aircraft engines. In: Hasic Telalovic, J., Kantardzic, M. (eds.) Mediterranean Forum – Data Science Conference. pp. 95–109. Springer International Publishing, Cham (2021)
2. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. CoRR **abs/1806.08049** (2018), http://arxiv.org/abs/1806.08049
3. Bobek, S., Bałaga, P., Nalepa, G.J.: Towards model-agnostic ensemble explanations. In: Computational Science – ICCS 2021, pp. 39–51. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-77970-2_4
4. Frank, A.G., Dalenogare, L.S., Ayala, N.F.: Industry 4.0 technologies: Implementation patterns in manufacturing companies. International Journal of Production Economics **210**, 15–26 (2019). https://doi.org/10.1016/j.ijpe.2019.01.004
5. Frederick, D., DeCastro, J., Litt, J.: User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS). NASA Technical Manuscript **2007–215026** (01 2007)
6. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**(3), 50–57 (Oct 2017). https://doi.org/10.1609/aimag.v38i3.2741
7. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (xai) program. AI Magazine **40**(2), 44–58 (Jun 2019). https://doi.org/10.1609/aimag.v40i2.2850
8. Hastie, T., Tibshirani, R.: Generalized Additive Models. Statistical Science **1**(3), 297 – 310 (1986). https://doi.org/10.1214/ss/1177013604
9. Khelif, R., Chebel-Morello, B., Malinowski, S., Laajili, E., Fnaiech, F., Zerhouni, N.: Direct remaining useful life estimation based on support vector regression. IEEE Transactions on Industrial Electronics **64**(3), 2276–2285 (2017). https://doi.org/10.1109/TIE.2016.2623260
10. Li, X., Ding, Q., Sun, J.Q.: Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety **172**, 1–11 (2018). https://doi.org/10.1016/j.ress.2017.11.021
11. Listou Ellefsen, A., Bjørlykhaug, E., Æsøy, V., Ushakov, S., Zhang, H.: Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. Reliability Engineering & System Safety **183**, 240–251 (2019). https://doi.org/10.1016/j.ress.2018.11.027
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40

S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
13. Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), https://christophm.github.io/interpretable-ml-book
14. Mosallam, A., Medjaher, K., Zerhouni, N.: Data-driven prognostic method based on bayesian approaches for direct remaining useful life prediction. Journal of Intelligent Manufacturing **27**, 1–12 (06 2014). https://doi.org/10.1007/s10845-014-0933-4
15. Nori, H., Jenkins, S., Koch, P., Caruana, R.: InterpretML: A unified framework for machine learning interpretability. CoRR **abs/1909.09223** (2019), http://arxiv.org/abs/1909.09223
16. de Oliveira da Costa, P.R., Akçay, A., Zhang, Y., Kaymak, U.: Remaining useful lifetime prediction via deep domain adaptation. Reliability Engineering & System Safety **195**, 106682 (2020). https://doi.org/10.1016/j.ress.2019.106682
17. Petkovic, D., Altman, R., Wong, M., Vigil, A.: Improving the explainability of random forest classifier – user centered approach. In: Biocomputing 2018. pp. 204–215. WORLD SCIENTIFIC (2017). https://doi.org/10.1142/9789813235533_0019
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778
19. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018)
20. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (May 2019). https://doi.org/10.1038/s42256-019-0048-x
21. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 International Conference on Prognostics and Health Management. pp. 1–9 (2008). https://doi.org/10.1109/PHM.2008.4711414
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019). https://doi.org/10.1007/s11263-019-01228-7
23. Wang, Y., Zhao, Y., Addepalli, S.: Remaining useful life prediction using deep learning approaches: A review. Procedia Manufacturing **49**, 81–88 (2020). https://doi.org/10.1016/j.promfg.2020.06.015, proceedings of the 8th International Conference on Through-Life Engineering Services – TESConf 2019

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_40