Explanation-driven model stacking

Szymon Bobek[1][0000−0002−6350−8405], Maciej Mozolewski[2], and Grzegorz J. Nalepa[1][0000−0002−8182−4225]

[1] Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and Institute of Applied Computer Science, Jagiellonian University, 31-007 Kraków, Poland
{szymon.bobek,grzegorz.j.nalepa}@uj.edu.pl
[2] Edrone Ltd.
m.mozolewski@edrone.me

Abstract. With advances of artificial intelligence (AI), there is a growing need for provisioning of transparency and accountability to AI systems. These properties can be achieved with eXplainable AI (XAI) methods, extensively developed over the last few years with relation for machine learning (ML) models. However, the practical usage of XAI is limited nowadays in most of the cases to the feature engineering phase of the data mining (DM) process. We argue that explainability as a property of a system should be used along with other quality metrics such as accuracy, precision, recall in order to deliver better AI models. In this paper we present a method that allows for weighted ML model stacking and demonstrates its practical use in an illustrative example.

Keywords: explainability · machine learning · optimization.

1 Introduction

Recent advancements in black-box machine learning models such as deep neural networks and their applications to sensitive areas such as medical and law applications, or industry 4.0 provoked a discussion on the accountability and transparency of AI systems [3]. Although the concept of explanation of decisions of AI systems has long tradition that dates back to times of knowledge-based AI systems [16], it has been extensively developed over the last decade to facilitate new algorithms.

A large portfolio of XAI algorithms is now available for data scientists and system engineers, which includes model-agnostic methods such as Lime [12], Shap [10], Anchor [13], and model specific solutions like GradCAM or DeepLift for neural networks [17]. However, the methodology of incorporating them into the classic data mining and machine learning pipeline is not clearly stated. In this paper we argue that the explainability (or intelligibility) is a property of a system as a whole and should be considered as an important factor in designing and evaluating such a system. This requires the explanation to be quantified with respect to some criterion.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

In this paper we show how the quality of explanation and the quality of machine learning model can be fused in order to produce a model that combines property of both, resulting in a model of good quality with explanations of good quality. In order to measure the quality of explanations we used *InXAI framework*[3] we developed, that provides objective metrics such as consistency, stability and perturbational accuracy loss [18]. We show how to combine these measures along with standard metrics for ML model performance (i.e. accuracy, F1, precision, recall, etc.) within a Bayesian optimization framework based on the SMAC toolkit [7] that stacks multiple ML models into one meta-model. We demonstrate the feasibility of our solution in an illustrative example.

The rest of the paper is organized as follows. In Section 2 we discuss the role of XAI methods in standard ML/DM pipeline and its potential usage as a criteria for model selection. Formal definition of our approach for explanation-driven model stacking is given in Section 3. In Section 4 we demonstrate the usage of our approach on illustrative, reproducible example. Finally in Section 5 we summarize the original contribution and indicate future works.

2 Role of XAI in the machine learning pipeline

Explainable AI aims at bringing transparency to the decision making process of automated systems. Along with the development of deep neural networks and other black-box machine learning methods, it has been extensively developed over the last decade. Both the recent GDPR EU regulation [6] and the DARPA-BAA-16-53 program on XAI [4] catalysed the progress in this field.

Although the general concept of explainable decision making is clear, the underlying methods and specific goals differ depending on who is the addressee of the explanation. Similarly, the location of the explanation mechanism in the pipeline of developing AI systems will be different depending on the end-user. In GDPR and DARPA documents the role of the end-user is emphasised, as the final recipient of the explanation. In such a case the *explainability* will be considered more of the property of an AI system as a whole and can be defined as a capability of the system to be understood. In the history of AI systems such a property was most often called intelligibility [9]. It was provided by building systems with frameworks that supported that feature inherently [5, 1, 15]. Nowadays it is addressed also by dedicated methods such as conversational recommender systems [8]. However, such approaches are crafted for the purpose of the specific problem and do not generalize well to other cases.

Most recent advancements in XAI focus mostly on generating explanation in a way that is mostly used by data scientists and domain experts to validate the correctness of the decision model (e.g. bias analysis), or to enable the adoption of the decision support systems in sensitive areas by building trust via explanations (e.g. medical diagnosis decision support systems). In both of the cases evaluation is done manually either via user-experience studies or by observational studies.

[3] See: https://github.com/sbobek/inxai.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

This is why it is difficult to incorporate the XAI methods within machine learning and data mining pipelines, which is a highly automated process.

To address these challenges, several approaches for automated evaluation of XAI methods were proposed. There were attempts to provide methodological approach for evaluation and verification of explanation results [11, 18]. Among many qualitative approaches there are also ones that allow for quantitative evaluation. In [14] measures such as fidelity, consistency and stability were coined, that can be used for a numerical comparison of methods. In [19] the aforementioned measures were used to improve overall explanations. In [2] a measure that allows us to capture stability or robustness of explanations was introduced. However, all of the solution provide only human-based evaluation procedure that does not produce objectively comparative results among different explanaibility methods. This limits their usage to use cases where expert-based analysis is the only one desired, discarding the possibility of including them in an automated pipeline and using their results as optimization parameters.

Our *InXAI framework* implements several metrics from aforementioned works, and allows include them in classic machine learning pipeline that is consistent with scikit-learn API[4]. This opens a possibility to use XAI metrics as any other machine learning model performance indicators and use them as model-selection attributes. In the next section we demonstrate how to use it along with Bayesian optimization framework to stack several machine learning models that finally yields high accuracy and good explainability properties.

3 Optimization of the explanation-driven meta model

Model selection is an important stage in building any AI system. Usually it is governed by the mechanisms that are based on comparison of standard metrics for machine learning models such as accuracy for classification or R2 score for regression. In this section we demonstrate how additional metrics associated with explainability can be combined with standard measures in order to facilitate model selection, but also with the model stacking mechanism.

Let us consider an example of a simple binary classifier. One can train several classifiers. These models will vary in terms of performance metrics, such as Accuracy, Logarithmic Loss, F1 Score to name a few. Depending on a specific data mining problem, it may also be important to take into account explainability.

Instead of choosing only one of the models, several "component/unit" models can be combined altogether to obtain a meta-model, so that specific performance metric remains at a decent level. At the same time, one may want to optimize XAI metrics, such as Stability, Perturbational Accuracy Loss or Consistency of the meta-model, emphasizing a specific aspect of explainability. The simplest way to obtain a meta-model for binary classifiers will be a weighted sum of k component models. Training of unit models is done independently. Suppose that the model predicts class 0 or 1. Each of those models predicts the probability of

[4] See https://https://scikit-learn.org.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

an instance $x^{(i)}$ belonging to a given class Q denoted as $P_k(Q|x^{(i)})$. Prediction P_{mm} of meta-model can therefore be defined as a weighted sum of predictions probability of its components and is given by Eq. (1).

$$\mathbb{P}_{mm}(Q|x^{(i)}) = \frac{\sum_k \mathbb{P}_k(Q|x^{(i)})w_k}{\sum_k w_k} \tag{1}$$

$$\sum_k w_k > 0;\ w_k \geq 0$$

On such a meta-model, result of the classification for observation $x^{(i)}$ is straightforward and can be defined as $\underset{Q}{\text{argmax}}\ \mathbb{P}_{mm}(Q|x^{(i)})$.

Where w_k is a weight associated with the model k and reflects the importance of that model in calculating global prediction. Such weight can be calculated as shares in the quality metric of a particular model (e.g. accuracy). In the following sections we show how w_k can be determined with the usage of XAI quality metrics and SMAC optimization framework.

3.1 Metrics of explainability

Meta model defined in Eq. (1) can be considered as a black-box mechanism, and easily used along with any ML quality metrics and XAI quality metrics. In the following section we briefly discuss selected XAI metrics that are used in our solution to measure the overall model performance in terms of quality of explanations.

Stability. It expresses to what extent are the explanations for similar observations similar to each other. This metric is specific for a single model. It is based on Local Lipschitz Continuity metric [2].

AUC Perturbational Accuracy Loss. This metric describes how accuracy metric changes along with increasing disruptions in predictor values. Like stability, this metric is defined for a single model. Perturbations are expressed as a percentage of random changes made to the data set. The smaller the weight of a variable given by a local explainer model, the larger perturbations are applied. The significance of the feature can be determined by permutation importance method (e.g. *Permutation Importance* from ELI5 package[5]).

Meta-model inner consistency. Consistency answers the question of how similar are the explanations of two or more different ML models that were trained on the same data set. It is a measure obtained on set o explanations $\{\Phi^{e_{m_1}}, \ldots, \Phi^{e_{m_k}}\}$ generated for k models and is defined by Eq. (2).

$$C(\Phi^{m_1}, \Phi^{m_2}, \ldots, \Phi^{m_k}) = \frac{1}{\max\limits_{a,b \in 1,2,\ldots,k} ||\Phi^{m_a} - \Phi^{m_b}||_2 + 1} \tag{2}$$

[5] See: https://eli5.readthedocs.io/en/latest/blackbox

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

The measure is applicable when one compares two or more (with the InXAI framework extension) different models. However, for the sake of clarity we limit the discussion to a single meta-model. Thus, we propose the *Inner meta-model consistency* measure given by Eq. (3). Note that Φ^{m_k} is an explanation generated with any explainer (eg. SHAP, LIME) for model k and is a matrix of i rows and n columns reflecting number of observations and number of features in a dataset respectively. Therefore the consistency of a meta-model C_{mm} is a vector of i elements.

$$C_{mm} = C\left(\frac{w_1}{\sum_k w_k}\Phi^{m_1}, \frac{w_2}{\sum_k w_k}\Phi^{m_2}, \ldots, \frac{w_1}{\sum_k w_k}\Phi^{m_k}\right) \tag{3}$$

For a meta-model constructed as a weighted sum of unit models, performance metrics will have lower bound equal to the metric of the weakest component model in the regard of the given metric. Therefore, by optimizing the weights of the meta-model components in terms of the selected XAI metric, one can be sure that the performance metrics for the resulting model will not fall below the expected level. The level is determined on the basis of the initial selection of the component models. This can be demonstrated on the example of the area under the ROC curve metric. Consider the first model m_1, which predicts the class on the basis of a random throw of an unbiased die ($P(0) = 0$ or $P(0) = 1$ with equal probability). For balanced classes, the area under the ROC curve for m_1 will be equal to 0.5. Let's assume that the second model m_2 will have an area under the ROC curve greater than 0.5. Let m_1 be included in the meta-model with weight w_1, and m_2 has a corresponding weight w_2. Then the area under the ROC curve of the meta-model will be greater than 0.5, as long as $w_2 > 0$.

3.2 Selection of weights of unit models in meta-model

To combine XAI metrics, the formula for Loss function L_{mm} for meta-model, allowing to put more emphasis on a given metric, depending on the course of the experiment, was developed. To control the extent to which a given XAI metric is important for optimisation, importance meta-parameters were introduced. For AUC Perturbational Accuracy Loss of meta-model ($AUCx_{mm}$), the γ_{auc} parameter was used. For stability (S_{mm}) γ_s and for consistency (C_{mm}) γc were used, respectively. The idea behind those parameters is to put more emphasis on the given metric by taking a given metric to the power of the parameter > 1. Stability and consistency are vectors, thus mean value across all observations were used. For details see Eq. (4).

$$L_{mm} = \frac{{AUCx_{mm}}^{\gamma_{auc}}}{\overline{S_{mm}}^{\gamma_s} \cdot \overline{C_{mm}}^{\gamma_c}} \tag{4}$$

Where $\overline{S_{mm}} = \frac{\sum_i^N {S_{mm}}^i}{N}$ and $\overline{C_{mm}} = \frac{\sum_i^N {C_{mm}}^i}{N}$ are defined as average stability and consistency on the dataset for selected models.

The next section provides an evaluation scenario of the framework using an illustrative example.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

4 Evaluation on a case study

In this section we demonstrate how the formal representation of framework given in Section 3 can be operationalized and enclosed into a working module. For the sake of clarity we demonstrate the solution on a synthetic, reproducible example.[6]

4.1 Synthetic dataset and ML models

For the purpose of a demonstration, a synthetic example was used. Dataset with two interleaving half circles was generated with the *sklearn* library. It contains 200 observations and is visualised on Fig. 1. There are two predictor variables. The test set consisted of 33% of the observations. Models trained on the dataset are summarized in Tab. 1.

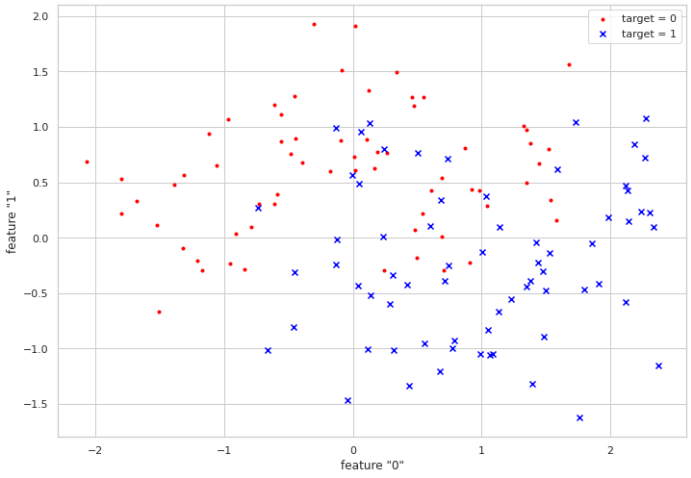

Fig. 1. Dataset with 2 classes and 2 predictors.

4.2 XAI metrics on unit models

Metrics were obtained according to methodology presented in Sect. 3.1 with the use of the *InXAI* framework. SHAP values were used as a local explainer. Stability per unit model is presented on Fig. 2. The highest stability characterizes SVM Classifier models, followed by RandomForest, with XGBoost in the middle and CatBoost as least performant. Consistency between models was calculated pairwise and is depicted in Fig. 3. High level of Consistency was shown only

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

model	model abbreviation	accuracy score	F1-score class "0"	F1-score class "1"
SVMClassifier with RBF kernel	svc_radial	0.76	0.78	0.73
SVMClassifier with linear kernel	svc_lin	0.82	0.83	0.80
XGBClassifier	xgbc	0.74	0.76	0.72
RandomForestClassifier	rfc	0.74	0.77	0.70
CatBoostClassifier	ctbc	0.65	0.66	0.65

Table 1. Summary of trained models.

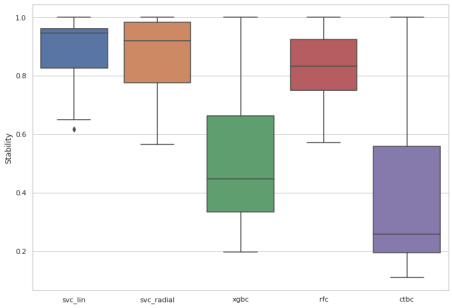

Fig. 2. Stability per unit model.

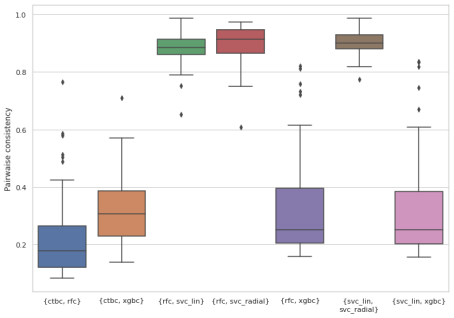

Fig. 3. Pairwise consistency.

between RandomForest and SVM models. In terms of AUC (Area Under Curve) for Perturbational Accuracy Loss, most models performed equally, with the one exception of CatBoost. This model was also the weakest in terms of initial – not perturbed – Accuracy. The summary is presented on Fig. 4.

4.3 Optimization-driven weight selection

It is worth noting that computation of XAI metrics is very computational intensive operation that does not scale well to large datasets. This is why optimization of parameters that are based on such metrics has to be performed wisely. In order to assure the feasibility of the approach the *SMAC* framework[7] was used for optimization of the Loss function L (finding the minimum) that is based on the model-based Bayesian optimization algorithm.

The first optimization run began with user-defined weights for the component models. Initial weights for component models were the same in all experiments. They were computed as follows:

$$w_k = \frac{1}{\gamma_{auc}^k}$$

[6] For source code see: https://github.com/mozo64/inxai/blob/main/examples/xai_on_synth_data/XAI-boost-on-syntetic-data-v4.ipynb.
[7] See: https://automl.github.io/.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

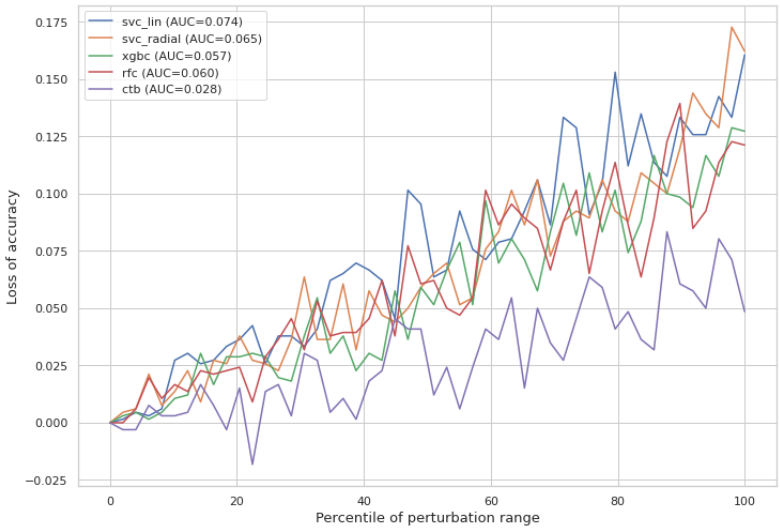

Fig. 4. AUC Perturbational Accuracy Loss.

and afterwards re-scaled to sum up to 100%. In the course of the experiment, the optimizer created a meta-model linking individual unit models according to Eq. (1). SMAC optimised the weights for all 5 unit models with weights in the uniform range from 0.0 to 1.0 and scale them so that they sum up to 100%. The Loss function L given by the formula Eq. (4) was minimized with different values of meta-parameters. 30 iterations per experiment were performed.

To speed up computations, two of the metrics – Stability and AUC acc. loss – were approximated with Eq. (5) and Eq. (6), respectively. However, this approximation was used only during the SMAC optimization process. The final result of the experiment was calculated with obtained weights for component models, using the exact equations from the Sect. 3.1.

$$S_{approx} = \frac{\sum_k S_k \, w_k}{\sum_k \, w_k} \tag{5}$$

$$AUCx_{approx} = \frac{\sum_k AUCx_k \, w_k}{\sum_k \, w_k} \tag{6}$$

4.4 Results

The summary of experiment runs is depicted in Tab. 2. In the run #3a the impact on Stability was comparable to run #4a (where we optimised for Inner meta-model consistency). Thus importance of Stability was further increased in run #3b. Likewise, in the run #4a the result for the Inner meta-model consistency was the same as in run #3a. So we increased the importance of the Inner meta-model consistency even further. Overall one can see impact on meta-model XAI

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

metrics in terms of AUC acc. loss, Stability and Consistency, what was the aim of the study. Meta-model Accuracy stayed on a decent level, as expected.

#	meta-parameter			weights for models after optimization					metrics			
	AUC acc. loss	Stabi-lity	Consis-tency	xgbc	rfc	ctbc	svc_lin	svc_radial	model acc.	AUC acc. loss	Stabi-lity	Consis-tency
1	1.0	1.0	1.0	.000087	.363524	.000031	.272740	.363619	0.76	0.060	0.872	0.895
2	**3.0**	1.0	1.0	.000042	**.499893**	.000025	**.000004**	**.500035**	0.73	**0.048**	0.858	0.862
3a	1.0	**3.0**	1.0	.000007	.315697	.000021	.312844	.371430	0.77	0.062	**0.874**	0.899
3b	1.0	**5.0**	1.0	.000021	**.000013**	.000020	**.499952**	**.499993**	0.77	0.059	**0.887**	0.871
4a	1.0	1.0	**3.0**	.000062	.318573	.000074	.310580	.370711	0.77	0.064	0.874	**0.899**
4b	1.0	1.0	**5.0**	.000026	.293124	.000037	**.350562**	**.356252**	0.77	0.067	0.876	**0.902**

Table 2. Results of experiment runs.

XGBoost and CatBoost Classifier models in all runs got low, negligible weights in the meta-model. It is related to low Stability of those models, in comparison to other models (see Fig. 2). $Lossfunction$ penalised low Stability in all experiment runs. Optimisation for AUC acc. loss (#2) resulted in the lowest weight for SVM Classifier with Linear kernel. This classifier has the worst performance in terms of AUC acc. loss – see: Fig. 4. Best Stability (#3b) was when the lowest weight was applied to RandomForest Classifier. This model has slightly worse Stability than both SVM Classifiers (see: Fig. 2). The best result when optimising for Inner meta-model consistency (#4b) was for SVM Classifier models (Linear kernel, RBF kernel) having similar weights. This is in line with the fact that those 2 models have highest pairwise consistency (see: Fig. 3).

5 Summary and future works

In this paper we presented a method that allows us to combine XAI quality metrics along with standard ML evaluation metrics in order to provide an optimization framework that maximizes both ML performance and XAI quality within a single meta-model. We demonstrated that in our approach there is no need to compromise on performance metrics, such as accuracy, as the meta-model preserves the quality of its components.

Also the *naive* approximation of Stability and AUC Perturbational Accuracy Loss for meta-model was given. This approximation is more effective in terms of CPU needed for computations. We also introduced the concept of the *inner meta-model consistency*, which shows its usefulness, as it promotes higher weights for models which were pairwise consistent.

The idea of an ensemble meta-model is worth further research. Firstly, especially valuable will be generalization on a multi-classifier problem and regression problem. In the future works, we can also check whether other XAI metrics than AUC Perturbational Accuracy Loss, Stability and Inner meta-model consistency

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

could be optimized in this framework. Secondly, it is also necessary to validate the method on real life examples. Finally, the idea of combining different explanations for one ML model into one meta-explanation is worth exploring. So instead of weighted sum of unit models, one could put together different local explainers (eg. SHAP, LIME), to create one meta-explainer, optimized for specific XAI metrics.

One of the limitations of the presented framework is that it only provides model weighting using comparative evaluation metrics among several models/explainers. It does not assure that the explanations generated by the explainer are correct nor feasible to the end user. It only takes into account their performance on the measurable, objective aspects. The fit to the expectations is another research topic that not yet has been operationalized in our framework.

Acknowledgements

The paper is funded from the PACMEL project funded by the National Science Centre, Poland under CHIST-ERA programme (NCN 2018/27/Z/ST6/03392). The authors are grateful to ACK Cyfronet, Krakow for granting access to the computing infrastructure built in the projects No. POIG.02.03.00-00-028/08 "PLATON - Science Services Platform" and No. POIG.02.03.00-00-110/13 "Deploying high-availability, critical services in Metropolitan Area Networks (MANHA)"

References

1. Almeida, A., Lopez-de Ipina, D.: Assessing ambiguity of context data in intelligent environments: Towards a more reliable context managing systems. Sensors **12**(4), 4934–4951 (2012), http://www.mdpi.com/1424-8220/12/4/4934
2. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods (2018)
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82 – 115 (2020)
4. DARPA: Broad agency announcement – explainable artificial intelligence (XAI). DARPA-BAA-16-53 (August 2016)
5. Dey, A.K.: Modeling and intelligibility in ambient environments. Journal of Ambient Intelligence and Smart Environments **1**(1), 57–62 (Jan 2009)
6. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a" right to explanation". arXiv preprint arXiv:1606.08813 (2016)
7. Hutter, F., Hoos, H.H., Leyton-Brown, K.: Sequential model-based optimization for general algorithm configuration (extended version). Tech. Rep. TR-2010-10, University of British Columbia, Department of Computer Science (2010), available online: http://www.cs.ubc.ca/~hutter/papers/10-TR-SMAC.pdf
8. Jannach, D., Manzoor, A., Cai, W., Chen, L.: A survey on conversational recommender systems (2020)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28

9. Lim, B.Y., Dey, A.K., Avrahami, D.: Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2119–2128. CHI '09, ACM, New York, NY, USA (2009). https://doi.org/10.1145/1518701.1519023, http://doi.acm.org/10.1145/1518701.1519023
10. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc. (2017)
11. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems (2020)
12. 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, https://doi.org/10.1145/2939672.2939778
13. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI Publications, Thirty-Second AAAI Conference on Artificial Intelligence (2018)
14. Robnik-Šikonja, M., Bohanec, M.: Perturbation-based explanations of prediction models. In: Zhou, J., Chen, F. (eds.) Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent, pp. 159–175. Springer (2018). https://doi.org/10.1007/978-3-319-90403-0_9, https://doi.org/10.1007/978-3-319-90403-0_9
15. Roy, N., Das, S.K., Julien, C.: Resource-optimized quality-assured ambiguous context mediation framework in pervasive environments. IEEE Trans. Mob. Comput. **11**(2), 218–229 (2012), http://dblp.uni-trier.de/db/journals/tmc/tmc11.html#RoyDJ12
16. Schank, R.C.: Explanation: A first pass. In: Kolodner, J.L., Riesbeck, C.K. (eds.) Experience, Memory, and Reasoning. pp. 139–165. Lawrence Erlbaum Associates, Hillsdale, NJ (1986)
17. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. CoRR **abs/1610.02391** (2016), http://arxiv.org/abs/1610.02391
18. Sokol, K., Flach, P.A.: Explainability fact sheets: A framework for systematic assessment of explainable approaches. CoRR **abs/1912.05100** (2019)
19. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D.I., Ravikumar, P.: On the (in)fidelity and sensitivity for explanations (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_28