A Study on Trust in Black Box Models and Post-Hoc Explanations

Nadia El Bekri[1], Jasmin Kling[1]and Marco F. Huber[2,3]

[1] Fraunhofer IOSB, Karlsruhe, Germany
nadia.elbekri@iosb.fraunhofer.de,
[2] Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany
[3] Center for Cyber Cognitive Intelligence (CCI), Fraunhofer IPA, Stuttgart, Germany

Abstract. Machine learning algorithms that construct complex prediction models are increasingly used for decision-making due to their high accuracy, *e.g.*, to decide whether a bank customer should receive a loan or not. Due to the complexity, the models are perceived as black boxes. One approach is to augment the models with post-hoc explainability. In this work, we evaluate three different explanation approaches based on the users' initial trust, the users' trust in the provided explanation, and the established trust in the black box by a within-subject design study.

Keywords: Machine Learning, Black Box, Explainability, Interpretability, Trust

1 Introduction

Decision-making based on machine learning is used in various applications to assist or even to replace human reasoning. Machines outperform humans when it comes to the amount of data they process in a short time. Furthermore, machines seem to make more consistent decisions since their only source for making decisions is the underlying data [2]. Nevertheless, if the data contains bias, the machines learn this as well. Since the models based on the machine learning algorithms are complex, the reasoning is mostly opaque. Trust has to be established [8] and is considered as one of the motivating aspects of intelligibility [7]. One approach to enhance intelligibility is to explain the decisions made by the complex model without entirely understanding how the model operates. The post-hoc explanations are generated by explainability methods on top of the black box. The human explanation approach is imitated by an explanation system, which can be used to provide explanations for these machine learning algorithms. An under-explored aspect thereby is to examine if the outcomes of explainability approaches really raise the users' trust in black box decisions.

The key contributions of this work is to examine if different explanation approaches really raise the users' trust based on the characteristics competency, reliability, consistency and understandability. We analyze how different explanation approaches are perceived and moreover examine if explanations are trusted by users and which aspects of the explanation are desirable. A human-grounded evaluation [4] of different explanation approaches to compare explanations based on the measured trust in the explanation and in the black box is performed.

2 Intelligibility and trust

Since intelligibility is a subjective concept, defining and measuring intelligibility is difficult. Doshi-Velez and Kim [4] characterize three evaluation approaches for intelligibility: application grounded, human grounded and functionality grounded. The application grounded uses the system understanding of a domain expert as the measure. The human grounded evaluation uses simpler experiments performed by an end user, not by a domain expert. The functionality grounded evaluation considers a formal measure.

2.1 Human subject studies and trust measures

Trust can be defined as "acceptance of an advice or an action, based on sufficient confidence in a positive outcome of that advice or action" [1]. For example, if domain knowledge already exists for a given task, the outcome of a prediction model can be seen as positive when the decision is in accordance to the domain knowledge. Based on this, trust can be measured as the ratio of the number of accepted actions and the number of all actions. It could be the case that faulty actions were accepted. This is why justifiable trust, the ratio of the number of accepted actions that had a positive outcome and the number of all actions, is desired. Poursabzi *et al.* [9] measure trust by determining the difference between the prediction of the model and the participant's prediction. As a use case, they predict housing prices. The users were given some information about the underlying prediction model and are asked to make a prediction of the housing price on their own. This estimation is compared to the housing prices predicted by the model. Their absolute deviation is a measure of trust whereas smaller values indicate higher trust. Ribeiro *et al.* [10] conducted a user study to measure if the participants will trust the prediction model. At first, the participants are given ten different instances and their black box predictions. Eight out of ten predictions were made correctly by the black box, whereas two instances were misclassified. The participants were asked to answer the following three questions: *"Do you trust this algorithm to work well in the real world? Why? How do you think the algorithm is able to distinguish between the classes?"* Next, the participants were given ten different instances together with their explanations and asked the same questions. These two sets of answers are evaluated against each other to determine whether providing an explanation increases the understanding of the aspects the underlying model uses to make a decision and whether this knowledge influences trust. As a result, it is stated that providing an explanation helps understanding the aspects an model uses for decision making and knowing when to trust it.

2.2 Post-hoc explanation approaches

Many different explanation approaches exist, *e.g.*, providing logical statements [16], local models [14], deep explanations [15], rule extraction [6], feature importance [5], and feature tweaking [13]. For the user study, we focus on the post-hoc explanation approaches Local model-agnostic explanations (LIME) [10], CluReFI (extended version of LIME) and treeinterpreter [11]. We focused on those approaches due the fact that they are all based on feature importance and follow a clear concept. LIME and CluReFI both list the most influential features, either for the given instance (LIME) or for the cluster representative (CluReFI). Besides linear classifiers, decision trees are the most widely used classification techniques in real world applications. Therefore, we included treeinterpreter since it's intuitive and easy to understand.

LIME: The approach [10] provides instance explanations that exploit the proximity of the instances to be explained. Sparse linear models are learned for providing local explanations. A local LIME explanation visualizes the prediction probability of the black box and the decision-relevant features together with their importance (see Figure 3) and the features that provide evidence against it. LIME answers the question *"What information did the model use to make this decision or which not?"*.

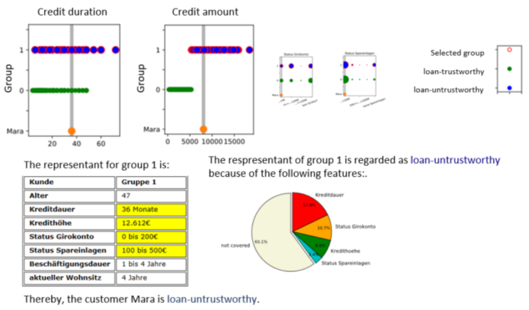
Cluster Representatives with LIME (CluReFI): In this paper, LIME was extended with additional information and is named CluReFI. The data was first clustered and the representative of the cluster was explained by LIME. The explanation (see Figure 1) first assigns an unseen data instance its closest cluster and visualizes this assignment using the range of validity per feature. CluReFI visualizes the feature validity ranges of each cluster for the most important features contributing towards the selected class (in Figure 1 four features are selected). For example, in Figure 1 on the top left the validity ranges for the features credit duration is displayed. The actual customer (not the representative) is the orange circle, additionally the group's loan-trustworthy (green) and loan-untrustworthy (blue) are displayed. The selected group for the customer is circled in red. This is done for all important features that contribute towards the selected class. On the bottom left of Figure 1, the tabular representation of the cluster representative is illustrated. The important features are highlighted to emphasize their importance. The third part of the explanation is the representatives explanation that illustrates a pie chart (see Figure 1 bottom right) of the most important features contributing towards the class of the representative and their importance. In contrast to LIME, CluReFI illustrates the user only the most important features contributing towards the class for the representative.

Treeinterpreter: The treeinterpreter interprets the predictions made by decision trees (DT) and random forests (RF) [11]. In contrast to LIME treeinterpreter is model-specific only for DTs and RFs. The trained model for the evaluation was a RF. Each instance prediction is decomposed into the prediction model's bias and the features contributing most to the instance prediction. The decision path is marked in red (see Figure 2).

3 Method

3.1 Participants

28 participants attended the study. The youngest participant was 19 years old and the oldest 58 years old. No specific background or requirements were demanded. Half of the participants were younger than 27. In total more men than women—79 percent compared to 21 percent—were surveyed. 43 percent of the participants had theoretical experience with machine learning, 14 percent had no experience at all, 7 percent had less than one year of practical experience, 25 percent had more than one year but



The representant for group 1 is:

Kunde	Gruppe 1
Alter	47
Kreditdauer	36 Monate
Kredithöhe	12.612€
Status Girokonto	0 bis 200€
Status Spareinlagen	100 bis 500€
Beschäftigungsdauer	1 bis 4 Jahre
aktueller Wohnsitz	4 Jahre

The respresentant of group 1 is regarded as loan-untrustworthy because of the following features:.

Thereby, the customer Mara is loan-untrustworthy.

Fig. 1. Explanation for CluReFI

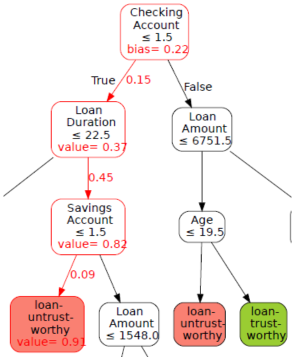
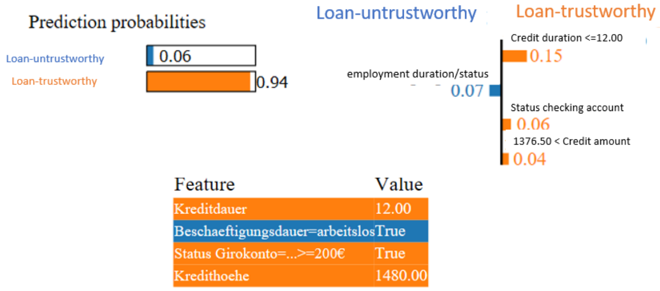


Fig. 2. Explanation for TreeInterpreter



Feature	Value
Kreditdauer	12.00
Beschaeftigungsdauer=arbeitslos	True
Status Girokonto=...>=200€	True
Kredithoehe	1480.00

Fig. 3. Explanation for Lime

less than 3 years practical experience, 4 percent had more than three years of practical experience and 7 percent specified another experience.

3.2 Materials

The user study was conducted via a graphical user interface [12] and was voluntary. No additional materials were allowed during the evaluation and no restrictions have been made for the users. Every explanation questionnaire was built up the same way (except the visualization of the specific explanation approach) to ensure the consistency.

3.3 Design

The user study is based upon a within-subject design. The setup of the user study was first evaluated in a pre-test before starting the main user study. Self-reports can be a critical source on a questionnaire but as the understanding of explanations is very subjective, we wanted to examine this additionally via self-reports. An interview with the participants in the pre-test phase was conducted, to examine if the measures were actually conform. Based on the German credit data set [3] regarding loan-trustworthiness, the use case is set up as follows: a financial institution is using a black box to classify a new customer as loan-trustworthy or loan-untrustworthy. Overall seven features (age, loan duration, loan amount, amount of checking account, amount of saving account, employment duration and the amount of years living in the present residence) of two exemplary customers are illustrated to the participants. Every customer was explained by the four approaches (three explanation approaches plus baseline explanation). The order of the approaches were randomized to counteract possible order effects and minimize learning across the conditions.

The two illustrated customers were the same for all participants. In the following, the hypotheses about the connection between explainability and trust, as well as the information needs are introduced.

Hypothesis 1: Participants with a bigger initial trust accept black box predictions more likely than participants with a lower initial trust.

Hypothesis 2: Providing an explanation increases the trust in the black box.

Hypothesis 3: Trust in an explanation correlates with the trust in the underlying black box.

Based on these hypotheses, different information needs have to be captured by the study. The variables regarding trust are divided into three categories: initial trust, trust in an explanation approach and trust in the underlying black box. In addition, the understandability of an explanation is measured. The four-point Likert scale was used to capture the participants' opinion on the information need except for the need for *understandability of an explanation*. The gradation of the scale were: *strongly agree, agree, disagree, strongly disagree*.

In the following, the information needs as well as the measurement are introduced and described.

Initial Trust (IT): First the participants' general trust (initial trust) is captured. This initial trust will serve as a base for categorizing participants based on their willingness and ability to trust. Therefore, statements regarding important decisions are listed and the participants are asked to specify their agreement with the given statements.

Understandability of an explanation (UE): The understandability can be measured by statements about the features that were considered by the explanation and the most important features used. For the understandability of an explanation each question has a certain number of possible answers per explanation. The number of features varies for every explanation approach and displayed customer. When only one choice can be selected, the participant receives one point for the correct answer and zero points for the wrong answer. Whenever multiple choices can be selected, the participant receives one point for each correctly selected choice, but looses one point per incorrectly selected choice. The more points a participant gains, the better he understands the explanation. The amount of possible points depends on the number of features that were used by the explanation approach. The points are scaled and converted to a score between 0 and 1.

Trust in an explanation (TE): Since self-reported understanding and actual understanding may differ, the explanation understandability metric is used to scale the self-reported understanding. When a participant does not understand the explanation, but reports to do so, his report is weakened based on his actual understanding. Trust in an explanation is queried based on eight statements (overall score between 0 and

24). This score is multiplied by the explanation understandability metric and then normalized to a range from 0 to 1. A final score of 1 indicates a very high trust in the explanation and a score of 0 indicates a very low trust in the explanation.

Trust in the black box (TBB): This metric measures the participants' trust in the underlying black box. The participants are queried about their understanding, reliability, consistency and competence of the black box model. The trust in the black box is queried via seven statements and rated on the four point Likert scale, an overall score between 0 and 21 can be reached. This score is scaled down to a range from 0 to 1. A final score of 1 indicates a high trust in the black box and 0 a low trust. The initial trust per participant is measured only once at the beginning of the user study and is grouped into the levels low, medium and high initial trust. The understandability of an explanation, the trust in an explanation, and the trust in the black box are measured once per explanation.

3.4 Procedure

At the beginning, the participants are informed about the topic, length and structure of the user study. Next, the different explanation approaches are introduced in a randomized order and evaluated by each participant. The first explanation only illustrates the prediction outcome (certain class) as well as the prediction probability. This explanation serves as a baseline when evaluating the explanations based on their understandability, trust in the explanation and trust in the black box that is generated by the explanation. Each presentation of the explanation approach is structured the same way. First, the contributions that the explanation approach makes are listed. Second, the explanation approach is described in more detail. Third, the explanation for the first customer is given and the participants understanding of the explanation is measured. Fourth, the explanation for the second customer is given and the participants understanding of the explanation is measured. Last, questions are asked about the explanations explainability and trust, as well as the trust in the black box model that is created by providing the explanations. Thereby, each participant evaluates eight explanations regarding the two exemplary customers.

4 Results

4.1 Trust variables

Based on the conducted user study and the aforementioned calculation of the different trust metrics, three different categories of trust variables exist: initial trust, trust in an explanation approach and trust in the underlying black box (see Appendix B for descriptive statistics of trust in the black box and trust in explanation). Since parametric tests require normality, the trust variables were tested for normality to ensure the applicability. Almost one third of the participants (32 percent) have a low initial trust that is characterized by an initialTrust value less than or equal to 0.2. Half of the participants have an initialTrust value greater than 0.2 but less than or equal to 0.4, belonging to the category medium initial Trust. The remaining 18 percent have an initial Trust value greater than 0.4 and are characterized by high initial trust.

Influence of demographic data: This part uses the collected data to examine if the demographics of the participants have an influence on the trust in black boxes and in explanations. The influence of the different demographic variables *age*, *gender*, and *experience* were tested with Analysis of variance (ANOVA).

Age: H_0 : No statistically significant relationship between the variable *age* and the variables *Trust*.

First, the trust in each explanation was tested. There was no significance for the variables *TEBaseline* with $F(3,24) = 1.6690,\ p = 0.2000$, *TELime* with $F(3,24) = 0.4380,\ p = 0.7280$, *TECluReFI* with $F(3,24) = 0.7800,\ p = 0.5170$, *TETree* with $F(3,24) = 0.1970,\ p = 0.8970$. Furthermore, the trust in the black box that is created by each explanation was tested. There was no significance for the variables *TBBBaseline* with $F(3,24) = 2.4710, p = 0.0862$, *TBBLime* with $F(3,24) = 1.295,\ p = 0.2990$, *TBBCluReFI* with $F(3,24) = 1.9150,\ p = 0.1540$, *TBBTree* with $F(3,24) = 0.3120,\ p = 0.8100$.

Gender: $H_0 : \mu_{\text{male}} = \mu_{\text{female}}$.

First, the trust in each explanation was tested. The variables for *TEBaseline* with $F(1,26) = 0.0780,\ p = 0.7830$, *TELime* with $F(1,26) = 2.1220,\ p = 0.1570$, *TECluReFI* with $F(1,26) = 0.1040,\ p = 0.7490$, *TETree* with $F(1,26) = 0.3130,\ p = 0.5810$ were insignificant. Furthermore, the trust in the black box that is created by each explanation was tested. There was no significance for the variables *TBBLime* with $F(1,26) = 0.1670,\ p = 0.6870$, *TBBCluReFI* with $F(1,26) = 0.0020,\ p = 0.9610$, *TBBTree* with $F(1,26) = 0.2900,\ p = 0.5950$. Only for *TBBBaseline* with $F(1,26) = 6.4190,\ p = 0.0177 < 0.05$ the results were significant.

Experience: $H_0 : \mu_{\text{no experience}} = \mu_{\text{theoretical experience}} = \mu_{<1 \text{ year practical experience}} = \mu_{\text{1-3 years practical experience}} = \mu_{>3 \text{ years practical experience}} = \mu_{\text{other experience}}$.

First, the trust in each explanation was tested. There was no significance for the variables *TEBaseline* with $F(5,22) = 0.9000,\ p = 0.4990$, *TELime* with $F(5,22) = 1.9760,\ p = 0.1220$, *TECluReFI* with $F(5,22) = 0.9100,\ p = 0.4920$, *TETree* with $F(5,22) = 1.0740,\ p = 0.4020$. Furthermore, the trust in the black box that is created by each explanation was tested. There was no significance for for the variables *TBBBaseline* with $F(5,22) = 0.9000,\ p = 0.4990$, *TBBLime* with $F(5,22) = 1.9760,\ p = 0.1220$, *TBBCluReFI* with $F(5,22) = 0.9100,\ p = 0.4920$. Only for *TBBTree* with $F(5,22) = 3.5250,\ p = 0.0172 < 0.05$ the results were significant.

Hypothesis 1: The hypothesis was tested with ANOVA. It is assumed that the different categories of the grouped variable *initialTrust* do not have an influence on the trust in black boxes (i) and in an explanation (ii).

$$H_0 : \mu_{\text{low}} = \mu_{\text{medium}} = \mu_{\text{high}}.$$

(i): At a confidence level of $\alpha = 0.01$, H_0 can only be rejected for the variable *TBBLime* with $F(2,25) = 6.0370,\ p = 0.0073$ (see results in Table 1). This indicates that the factors of initial trust interact very strong with the trust in black boxes generated by the explanation approach LIME. At a confidence level of $\alpha = 0.05$, H_0 can additional be rejected for the variable *TBBTree* with $F(2,25) = 4.5700,\ p = 0.0204$ (see results in Table 1).

Table 1. Results of Hypothesis 1 (i)

Approach	df	MSE	F	p-value
TBBBaseline				
Between Groups	2	0.0023	0.1660	0.8480
Within Groups	25	0.0761		
TBBLime				
Between Groups	2	0.1447	6.0370	0.0073
Within Groups	25	0.0240		
TBBCluReFI				
Between Groups	2	0.0543	1.1430	0.3350
Within Groups	25	0.0474		
TBBTree				
Between Groups	2	0.1220	4.5700	0.0204
Within Groups	25	0.0267		

Table 2. Results of Hypothesis 1 (ii)

Approach	df	MSE	F	p-value
TEBaseline				
Between Groups	2	0.0181	0.6920	0.5100
Within Groups	25	0.0262		
TELime				
Between Groups	2	0.1192	3.5620	0.0435
Within Groups	25	0.0335		
TECluReFI				
Between Groups	2	0.0910	1.8050	0.1850
Within Groups	25	0.0504		
TETree				
Between Groups	2	0.1860	5.0930	0.0140
Within Groups	25	0.0365		

(ii): At a confidence level of $\alpha = 0.05$, H_0 can be rejected for the variable *TELime* with $F(2, 25) = 3.5620,\ p = 0.0435$ and *TETree* with $F(2, 25) = 5.0930,\ p = 0.0140$ (see results in Table 2).

Furthermore, we used the Post-Hoc-Tukey test for *TBBLime*, *TBBTree*, *TELime* and *TETree* (see detailed results in Appendix C). There is a significant difference between the participants with medium initialTrust and high initialTrust.

For the LIME and the treeinterpreter explanation approach, it can be concluded that participants with a medium initial trust are more likely than other participants with a higher or lower level of initial trust to build trust in the underlying black box. Participants with a low or high level of initial trust are less likely influenced by providing an explanation. The assumption that participants with a higher initial trust accept black box predictions more likely than participants with a lower initial trust could not be confirmed for those variables. There are two explanations that are plausible for this behavior. Either, the initial trust of humans has no influence on their trust in black box predictions, or the collected data the variable *initialTrust* constitutes of is not actually measuring what is regarded as a humans' initial level of trust.

Hypothesis 2: It is assumed that the mean of the variable measuring the baseline for the trust in the black box and the means of the variables measuring trust in the black box which result from providing an explanation are equal.

$$H_0 : \mu_{\text{baseline}} = \mu_{\text{explanation approach}}.$$

Paired sample t-tests are conducted to test this hypothesis. All of the tests were significant at a significance level of $\alpha = 0.01$. Assuming that the mean of the baseline trust and the means of the trust generated by an explanation are equal does not hold. It can be concluded that the explanation increases the trust of the explanation receiver in the black box.

Hypothesis 3: We conducted three Pearson correlation tests to examine
H_0 : There is no significant relationship between the variable *TE* and the variable *TBB*.

The hypothesis can be rejected for each test: The association between the variable *TELime* and the variable *TBBLime* was significant with $r(26) = 3.7945,\ p =$

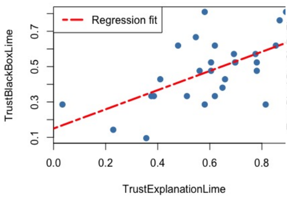


Fig. 4. y=0.5428x+0.1499, $R^2 = 0.3564$

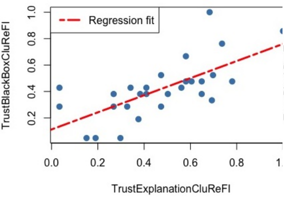


Fig. 5. y=0.6471+0.1127, $R^2 = 0.4663$

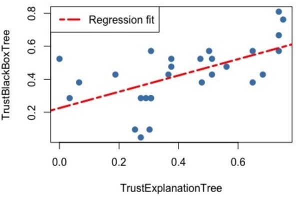


Fig. 6. y=0.4942x+0.2255, $R^2 = 0.3444$

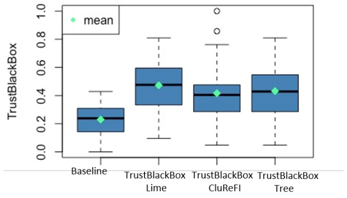


Fig. 7. TBB Scores

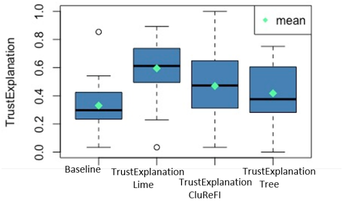


Fig. 8. TE Scores

$0.0008 < 0.01$. The pair *TECluReFI* and *TBBCluReFI* was significant with $r(26) = 4.7661,\ p = 0.0000 < 0.01$ and the variables *TETree* and *TBBTree* were significant with $r(26) = 3.6955,\ p = 0.0010 < 0.01$. The correlations for each explanation approach are illustrated in Figures 4, 5 and 6.

Best explanation: The best explanation is determined by the variables trust in the black box, trust in the explanation and the provided rankings. To examine the hypothesis ANOVA tests for TBB-Scores (see Figure 7), TE-Scores (see Figure 8) and Rankings with Post-Hoc-Tukey tests are performed. For the TBB vector the scores of TBBBaseline, TBBLime, TBBCluReFI and TBBTree were combined. Afterwards the vectors for the groups were created to illustrate the belonging of a group to an approach. That was equally done for the TE-Scores and the rankings. The results were significant with a significance level of $\alpha = 0.01$ for TBB with $F(3, 108) = 10.2400,\ p = 0.0000$, TE with $F(3, 108) = 8.2140,\ p = 0.0000$ and Ranking with $F(3, 108) = 48.5000,\ p = 0.0000$. For the Post-Hoc-Tukey tests pairwise comparisons were conducted. For TBB the pair (CluReFI (C)-Baseline (B)), (LIME (L)-B) and (Tree (T)-B) were significant with $\alpha = 0.01$. For TE the pair (L-B) and (T-L) were significant with $\alpha = 0.01$ and for Ranking the pair (C-B), (L-B), (T-B) and (L-C) were significant with $\alpha = 0.01$. In addition, the participants provided reasons for their ranking. LIME was preferred by 43% of the participants due to its understandability that is based on the familiarity with reading the explanation and a good compromise between level of detail and simplicity. The treeinterpreter was preferred by 39% of the participants due to its clear overview and good visualization. The representation is able to make the decision process, the steps it takes and the limits of the decision visible. Participants preferred CluReFI by 18% due to its informativeness regarding the cluster ranges. Participants felt to grasp the concept of the model and see if there exists a pattern.

5 Conclusion

A common thought is that explanations for black boxes definitely raise the users trust, but little research examines it. Although the prediction model and explanation are independent of each other, trusting the explanation approach is an important aspect of trusting the prediction model. The user study illustrates that creating explainability for the black box is valued by the recipient of the decision. The created explanation leads to a better understanding and thus acceptance of the black box. The user study indicated that for participants understandability and simplicity are essential. It is important to pick an optimal solution between informativeness and simplicity regarding a specific task. Our future work will consider a larger sample size and an improved design of the study design. Furthermore, we will consider participants with a high level of domain knowledge, which is likely to lead to even less trust in the early stages.

References

1. Alexandrov, N.: Explainable ai decisions for human-autonomy interactions. In: 17th AIAA Aviation Technology, Integration, and Operations Conference. p. 3991 (2017)
2. Davenport, T.H., Harris, J.G.: Automated decision making comes of age. MIT Sloan Management Review 46(4), 83 (2005)
3. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
4. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
5. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24(1), 44–65 (2015)
6. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154 (2017)
7. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
8. Mohseni, S., Ragan, E.D.: A human-grounded evaluation benchmark for local explanations of machine learning. arXiv preprint arXiv:1801.05075 (2018)
9. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability. arXiv preprint arXiv:1802.07810 (2018)
10. 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. pp. 1135–1144. ACM (2016)
11. Saabas, A.: treeinterpreter. https://github.com/andosa/treeinterpreter (2015)
12. SoSciSurvey: Soscisurvey tool. https://www.soscisurvey.de/ (2018)
13. Tolomei, G., Silvestri, F., Haines, A., Lalmas, M.: Interpretable predictions of tree-based ensembles via actionable feature tweaking. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 465–474. ACM (2017)
14. Turner, R.: A model explanation system. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1 – 6. Vietri sul Mare (2016)
15. Tzeng, F.Y., Ma, K.L.: Opening the black box-data driven visualization of neural networks. In: Visualization, 2005. VIS 05. IEEE. pp. 383–390. IEEE (2005)
16. Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., MacNeille, P.: Or's of and's for interpretable classification, with application to context-aware recommender systems. arXiv preprint arXiv:1504.07614 (2015)