Name: Felix Feldmann
Course: Seminar: Explainable Machine Learning
Student number: 3145215
Date: June 30, 2018

Seminar: Explainable Machine Learning

Measuring Machine Learning Model Interpretability

Seminar Report

Felix Feldmann

Contents

1 Introduction

Research in the field of machine learning systems has experienced a boost during the last decade. Nevertheless, new techniques like neural networks, becoming more sophisticated, where even experts do not always know why the underlying algorithm made a certain decision. Due to the lack of *interpretability* of machine learning algorithms and the demand of *explainablitiy* of the resulting decisions recent studies deal with the issue on how to make machine learning algorithms and their outcome *explainable*. The papers address the problem of what *interpretability* respectively *explainability* of an machine learning algorithm for a human means. To do so, two recent papers and their approaches will be presented, which cover this topic: First [7] by Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, and Wallach who tried to measure factors on interpretability e.g. trust in the model and simulatebility of a model to enhance the understanding of what is human-interpretable and secondly [6] by Narayanan, Chen, He, Kim, Gershman, and Doshi-Velez who ran a similar test, but focused on different factors e.g. the length of an explanation of an model, how the complexity of a model influences the trust in the model.

1.1 Motivation

The need for a explainability for machine learning algorithms is given by several reasons. First it is quite obvious, that anyone wants to understand why a certain decision was made and how different factors influenced the decision. Imagining the case of someone who wants to get a loan from a bank and an algorithm decides, it could be possible, that the underlying weights for the decision, discriminate the person e.g. because he/she lives in a poor suburb. Therefore it should be possible for the bank employee to *explain* the algorithms decision. Secondly, considering the above example, there will be a legal necessity in Europe, when the new GDPR (General Data Protection Regulation) is enforced in May 25. This law claims, that "[...] meaningful information about the logic involved [...]"[1] should be provided to explain machine decisions. Anyhow, it is not clearly defined, what *"meaningful information"* means. In order to understand, that the problem of interpretability is not an easy task to address we can imagine different users and their needs. Considering a CEO who wants to make better decisions for his company he needs different approaches to understand the algorithm then a data scientist who wants to debug is model. The following two publications address the measure-

[1] http://www.privacy-regulation.eu/en/article-15-right-of-access-by-the-data-subject-GDPR.htm, last seen June, 28th, 2018

ment of interpretability for a general case and deal solely with lay people, who do not have any particular deep knowledge of machine learning or data processing. The goal is to apply an approach to understand the fundamental properties of human behavior relevant to interpretability.

1.2 Related Work

First, Lakkaraju, Bach, and Leskovec [2] who proposes interpretable decision sets, which is a framework for building predictive models that are accurate and also interpretable. Decision sets are sets of independent if-then rules. Secondly, Mehrotra, Hendley, and Musolesi [5] created a machine learning app which suppresses unwanted mobile notifications. Therefore, the focus of the app was on the interpretability such that it is not a black box solution. Furthermore, Ribeiro, Singh, and Guestrin [8] introduced LIME, which is a explanation technique, that explains the predictions of any classifier in an interpretable way, by learning an interpretable model locally around the prediction. More publications regarding the topic of interpretability can be seen in [9] [1] [3] [4].

2 What is Interpretability?

To understand, what interpretability means, the two definitions of *interpretable* and *explain* by the Camebridge dictionary are given:

Definition Interpretable If something is interpretable, it is possible to find its meaning or possible to find a particular meaning in it: The research models failed to produce interpretable solutions.[2]

Definition Explain To make something clear or easy to understand by describing or giving information about it: If there's anything you don't understand, I'll be happy to explain.[3]

In the following the two terms *explainability* and *interpretability* are used equivalent. As it is quite challenging to give a good definition or description of what *Interpretability* for machine learning algorithms actually means, the two publications deal with interpretability as a latent problem, which cannot be measured directly. Moreover, the two

[2] https://dictionary.cambridge.org/us/dictionary/english/interpretable
[3] https://dictionary.cambridge.org/us/dictionary/english/explain

approaches measure the factors, which influence interpretabilty e.g. the machine learning model itself or the number of features it takes and also measure the factors which are influenced by interpretabilty e.g. trust or the ability to simulate a model. The key idea of the publications is to randomize the system factors on the input side and measure the impact of the human properties on the output side. Here its important to highlight, that not a specific machine learning approach is evaluated. The focus here is to evaluate the interpretability of machine learing output. In Figure 1 an overview of factors, which can influence interpretabilty and also the factors, which are influcenced by interpretabilty can be seen.

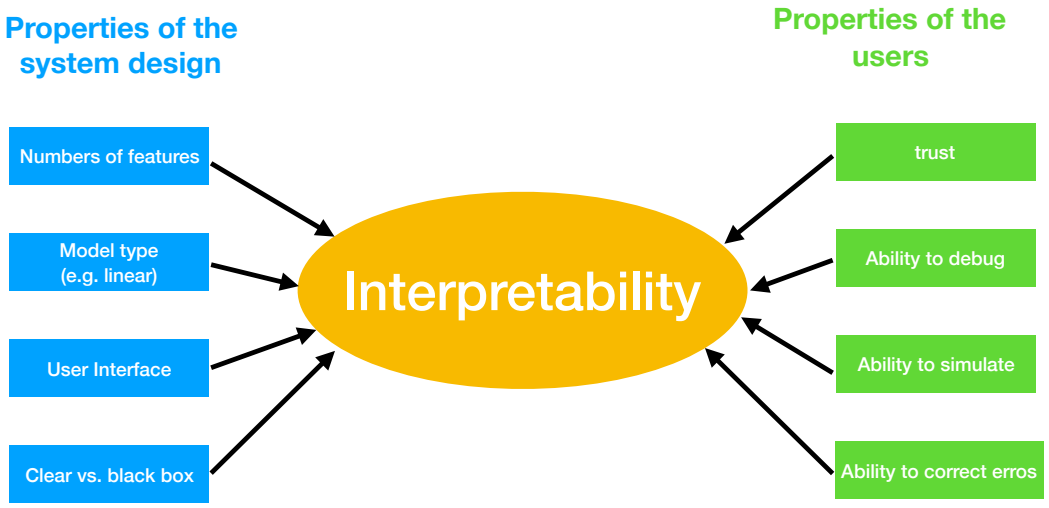


Figure 1: Interpretability as a latent problem. Left *Properties of the system design*. Factors which have an impact on *Interpretability*. Right *Properties of the users*.Factors, which are measurable outcomes of *Interpretability*.

2.1 Approaches for Interpretability

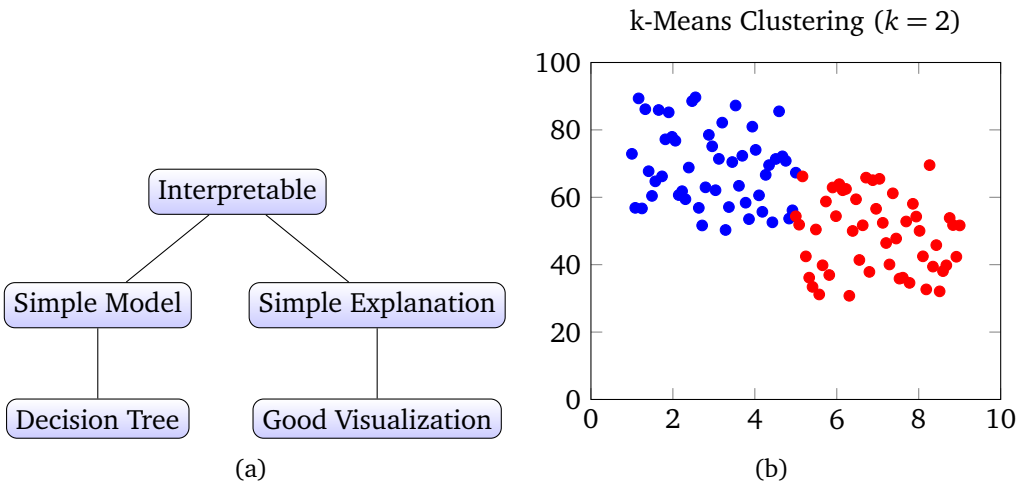
Interpretability for any kind of machine learning output has continuously been a topic, which was important for anyone to understand and comprehend the output, or even to validate, that it is right.

Simple Models

One of the easiest solutions to address the interpretability is to create and design simple models, which are easy to validate for the user. As an example could be small decision trees, where its easy to reproduce and reconstruct the decision of the algorithm (see Figure 2a).

Design of Simple Explanations

As models and the data, which has to be evaluated to make a decision, tend to get more and more complex it is important to have simple explanations. As a standard approach, the design of simple explanations e.g. the plotting of a colored clustering, can help to better understand why a certain point belongs to a class or not (see Figure 2b).



(a) (b)

Figure 2: Approaches to address interpretability. (a) Design of simple models e.g decision trees, where its easy to follow up a decision.(b) Creating easy explanations. Here k-Means clustering with a visualization of the clusters and its labels. In this case, it would be easy to understand, why a point is referred to a certain cluster.

3 Measuring Interpretability

Both publications ran experiments on interpretability, where users used to get identical experiments, only varying in factors related to interpretability. Nevertheless, both predefined similar, but slightly different goals:

- Apply approach to understand the fundamental properties of human behavior relevant to interpretability. [7]
- What kind of explanation are truly human interpretable and which are poorly understood?[6]

3.1 Manipulating and Measuring Model Interpretability

Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, and Wallach in [7] ran a randomized human subject experiment on 1250 participants from Mechanical Turk and varied factors

like the number of features or different types of models. Then they measured different outcome on different factors, e.g. the trust in the model, simulatability or the error of the end users prediction.

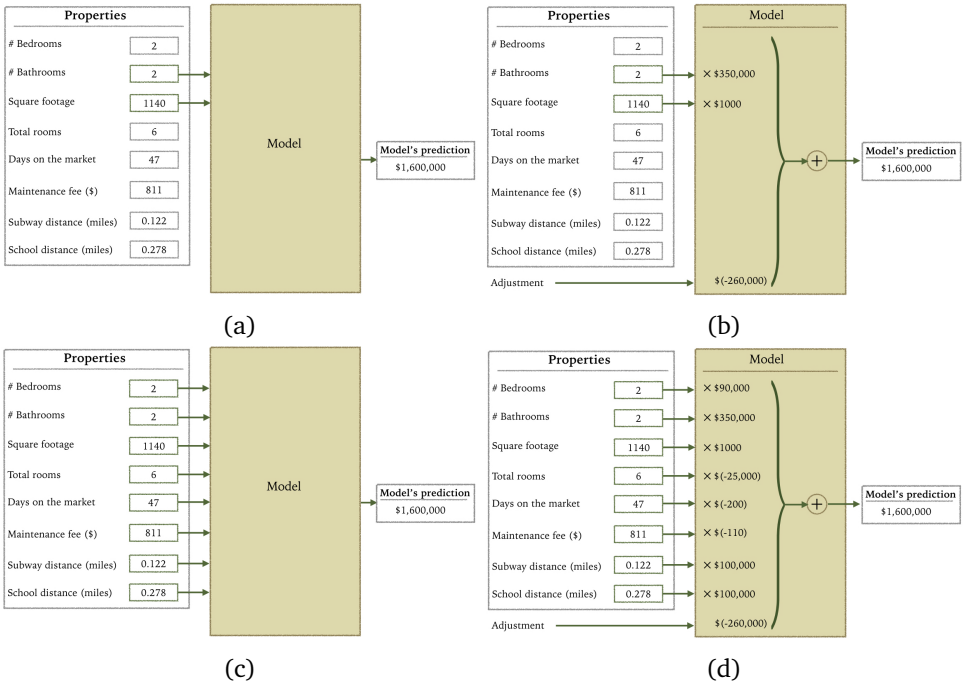


Figure 3: Experimental Conditions: Illustration of the four different models displayed to the users. (a) 2-features black-box model, where the internal weights of the model cannot be seen. (b) 2-features clear-box model, where the internal weights for the model's prediction can be seen by the user. (c) 8-features black-box model, internal weights are not displayed. (d) 8-features clear-box model, internal weights are displayed to the user.

In the task the users were asked to predict prices of apartments in New York city with the help of a model. Therefore, five different experimental conditions have been used (see Figure 3). The fifth condition, not mentioned in the figure, is the baseline, where the users had no help of a model. The differences between the models are, that they vary in the number of features used for predicting the prices and the type of the model, which means whether the insights of the model could be seen (clear-box model) or not (black-box model).

First the users entered a training phase, where they were shown 10 models and their corresponding predictions. Then the users entered a test phase. For each apartment the user were asked what the model will predict, in order to understand how good the participant understood the model (simulateability) and the user were asked about their

confidence in their prediction. Next, the actual prediction of the model has been shown and the user was asked what he/she thinks for how much the apartment actually was sold for. This step is used to check the users ability to correct errors in the prediction.

3.1.1 Results

While running the experiment the authors pre-registered[4] three hypotheses.

1. The clear, 2-feature model will be easiest for participants to simulate.
2. Participants will follow the clear, 2-feature model more than the black-box, 8-feature model.
3. Behavior will vary across conditions when an unusual example leads a model to make a highly inaccurate prediction.

For the first hypothesis we look at the *simulation error*, illustrated in Figure 4a, which is defined as follows: |model prediction − users guess of model prediction|. As hypothesized, the lower simulation error occures in the simpler clear-2-feature model. In addition it can be seen, that the clear-8-feature does have similar results in comparison to the black-box-models, which means, that transparency is not only relevant, also the number of features. For the second hypothesis the deviation and prediction error will be considered. The deviation error is defined as |model prediction − participants final prediction| and obviously the hypothesis does not hold (see Figure 4b). All models have the same impact on the peoples prediction. Taking a look at the prediction error in Figure 4c, which is defined as |actual price − participant's prediction|, no significant difference between the four models can be seen, but the model in general helps the users to predict slightly better, than without any model.

For the third hypothesis, the authors created apartments which had an unusual high number of bathrooms and therefore, the model predicted high prices for apartments with e.g. three bathrooms and only one extra room. The hope of the authors was, that participants who see the models internals spot the mistake and are able to make corrections. If people differ, when they see that the model is actually "bad", then we should see larger deviation for the clear models in Figure 4d. Counterintuitively no significant difference between the models could be observed, hence, the visibility of the features has no impact.

[4] Pre-registered experiments are used to create hypotheses before running the experiments to avoid fishing in the results till you find a corresponding hypotheses.

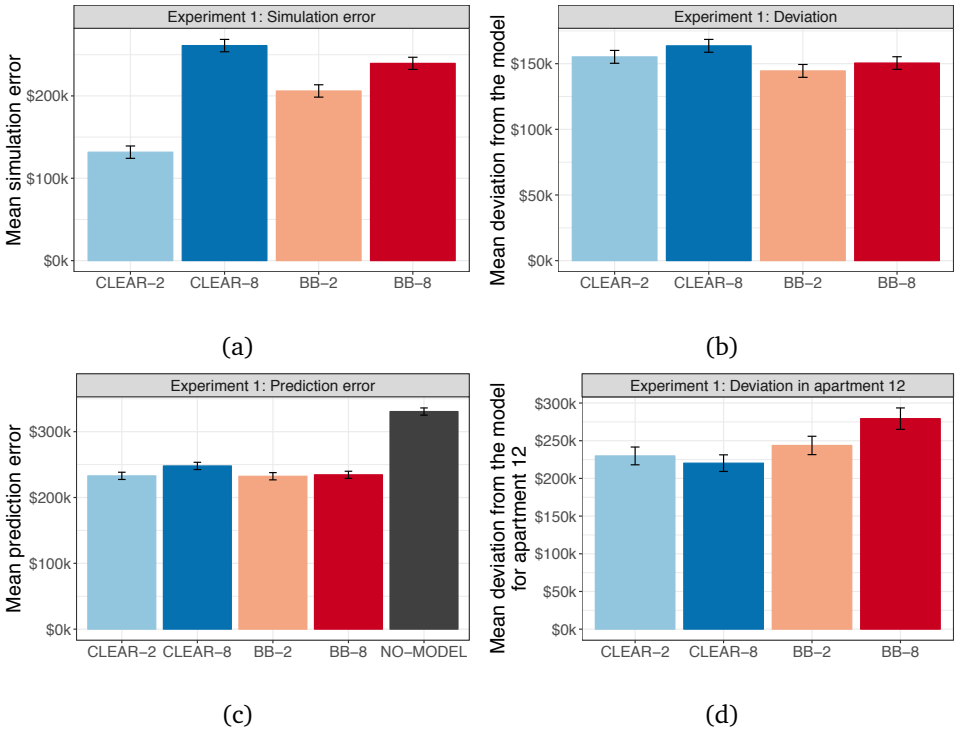


Figure 4: Results for the Housing predictions: (a) Simulation error of the users. (b) Deviation error. Deviation of the users from the models prediction. (c) Prediction error. The error users made from the actual price. No-model is the baseline model with the users not having any model. (d) Deviation error for a bad example. Here an artificial anomalous apartment with high number of bathrooms, leading to an extreme high price. Users who saw the internals did not spot the error.

Due to the exceptionally high prices for New York, the authors repeated the experiments by scaling down the prices of the apartments to a factor of 10. The underlying idea behind this scale-down was, that most people usually do not deal with such uncommon high prices and therefore the predictions would be biased. Anyhow, the experiments with the scaled down prices did not result in better predictions of the users.

3.2 How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation

In this experiment of Narayanan, Chen, He, Kim, Gershman, and Doshi-Velez in [6], the authors recruited 600 participants from Amazons Mechanical Turk. During the experiments different variations of the input parameters, e.g. variation of explanation size, new type of cognitive chunks, repeated terms in an explanation and domain variation

has been done. Additionally, measurements of the response time (time to click "Submit Answer"), accuracy and the subjective satisfaction have been taken.

In Figure 5, the experimental interface can be seen. The *alien's preference* can be seen as the model with its trained rules inside. Each row of the rules can be either implicit or explicit and the length (number of rows) has been varied in the experiments and are called *Cognitive Chunks*. *Observations* are the input to the machine learning algorithm, which has been varied in length. *Recommendation* corresponds to the models prediction. The *aliens preference* and the *Recommendation* are defined together as the *Explanation*.

The user were shown different inputs in two different domains, one in a medical domain, where the users should predict a medication for the alien and the other one in a food domain.

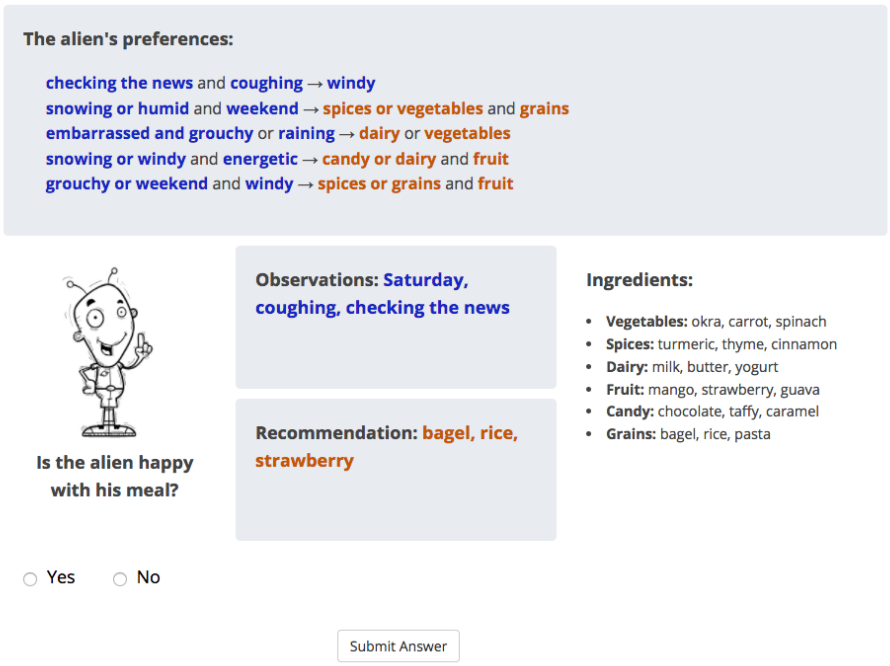


Figure 5: Experimental Conditions: *Aliens preferences* can be seen as the model itself. *Observations* as the input of the machine learning algorithm and *Recommendation* as the prediction of the model. *Aliens Preference and Recommendations* are defined as the *Explanation*.

3.2.1 Results

Same as with the previous paper, the authors hypothesized different hypothesis:

1. Increasing the size of the explanation either preferences or recommendations would increase the time to perform the task.

2. Adding cognitive chunks increases the time required to process an explanation.

3. If an input condition appeared in several lines of the explanation, it increases the time too find the correct rule.

4. Similar results for the clinical domain.

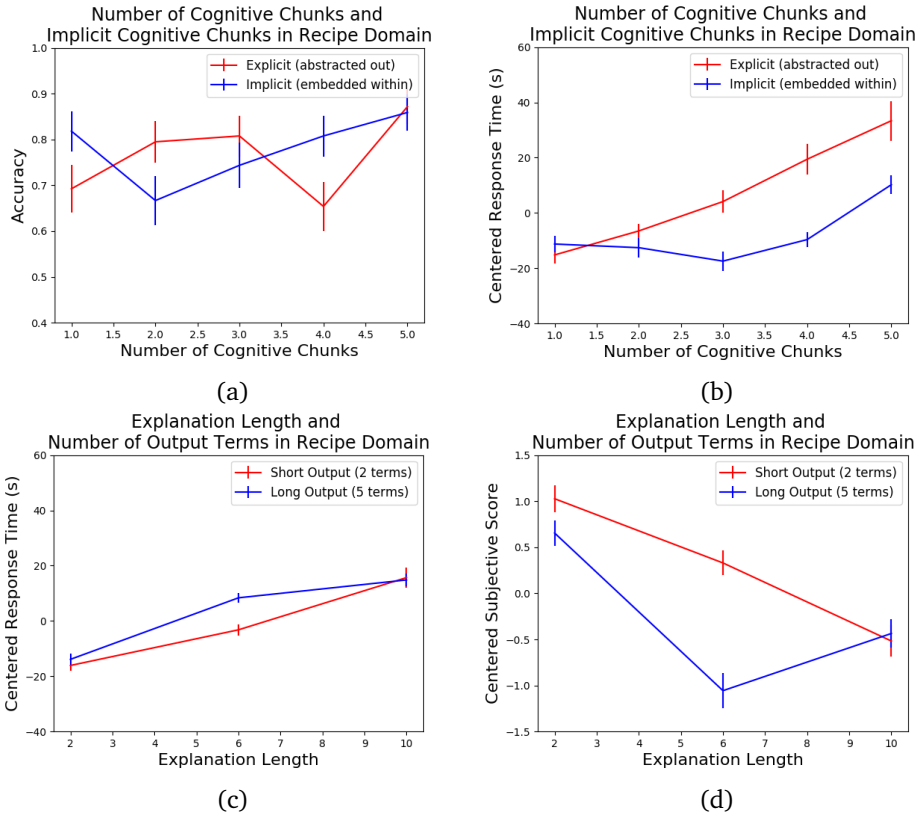


Figure 6: Experimental Conditions: Illustration of the four different models displayed to the users.

For the first hypotheses (see Figure 6c) the hypothesis holds, a longer explanation length clearly increases the time the user needs for a response. As well for the second hypothesis, the response time increases with the number of cognitive chunks. Interestingly the authors could observe, that explicit formulated chunks seem to be harder to process for the users then implicit ones. This also could be validated through the accuracy, which was measured during the survey and can be seen in Figure 6a, 6b. Therefore,

the accuracy increases constantly for implicit chunks, while for the explicit chunks its alternating. Differences between the clinical or the food domain could not be measured.

4 Summary & Conclusion

Both approaches addressed the question how to measure interpretability and to identify factors, which affect the ability to interpret machine learning models. Hereby, the focus was on lay people, with no specific background knowledge of machine learning or knowledge about the domain. The factors evaluated from the "system design" and the "users design" (see Figure 1)could have been more varied as well as the underlying model. Here the authors in both experiments tested only one model, therefore the question is which explanation or underlying model in what context is the best, e.g. decision trees or pseudocode. The main results of the two publications are, that for the interpretability, it does not really play a role to have a specific domain knowledge and counterintuitively the number of features seem to play a major role than if the model is a black- or a clear-box model. It could also be shown, that the length of explanation and therefore the time processing the output influences the perceived interpretability of a model.

Nevertheless, the topic recently emerged in 2017 also due to the GDPR, such that a vast potential in research lies ahead.

References

1. P. W. Koh and P. Liang. “Understanding black-box predictions via influence functions”. *arXiv preprint arXiv:1703.04730*, 2017.
2. H. Lakkaraju, S. H. Bach, and J. Leskovec. “Interpretable decision sets: A joint framework for description and prediction”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1675–1684.
3. Z. C. Lipton. “The mythos of model interpretability”. *arXiv preprint arXiv:1606.03490*, 2016.
4. Y. Lou, R. Caruana, and J. Gehrke. “Intelligible models for classification and regression”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, pp. 150–158.
5. A. Mehrotra, R. Hendley, and M. Musolesi. “Interpretable machine learning for mobile notification management: An overview of prefminer”. *GetMobile: Mobile Computing and Communications* 21:2, 2017, pp. 35–38.
6. M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez. “How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation”. *arXiv preprint arXiv:1802.00682*, 2018.
7. F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. “Manipulating and measuring model interpretability”. *arXiv preprint arXiv:1802.07810*, 2018.
8. M. T. Ribeiro, S. Singh, and C. Guestrin. “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*. ACM. 2016, pp. 1135–1144.
9. R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. “Grad-CAM: Why did you say that?” *arXiv preprint arXiv:1611.07450*, 2016.