Beyond the barrier of mistrust – an overview of selected methods for explaining predictions of machine learning models.

Filip Wójcik
Wroclaw University of Economics and Business
7 PUBLICATIONS 0 CITATIONS
SEE PROFILE

CREATIVETIME
doktorant.com.pl

18 ZAGADNIENIA AKTUALNIE PORUSZANE PRZEZ MŁODYCH NAUKOWCÓW

Redaktor Wydania: Krzysztof Piech
AGH Akademia Górniczo-Hutnicza, Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej, al. A. Mickiewicza 30, Kraków 30-059.

Skład tekstów i projekt graficzny okładki: Marcin Kuczera

Korekty: Krzysztof Piech, Marcin Kuczera

Opracowanie pt. ZAGADNIENIA AKTUALNIE PORUSZANE PRZEZ MŁODYCH NAUKOWCÓW 18 zawiera recenzowane prace naukowe Młodych Naukowców współpracujących z CreativeTime. Wybrane prace zostały nadesłane przez poszczególnych Uczestników Konferencji Młodych Naukowców nt. NOWE TRENDY W BADANIACH NAUKOWYCH - WYSTĄPIENIE MŁODEGO NAUKOWCA - I edycja – 20-21.06.2020 i II Edycja – 20-22.11.2020, Konferencji Młodych Naukowców nt. NOWE WYZWANIA DLA POLSKIEJ NAUKI - VII Edycja – 5-7.09.2020 oraz Konferencji Młodych Naukowców nt. ANALIZA ZAGADNIENIA, ANALIZA WYNIKÓW - WYSTĄPIENIE MŁODEGO NAUKOWCA – II edycja cd – 17-18.10.2020.

Skład opracowania wykonano na podstawie dostarczonych przez autorów tekstów. Wszystkie artykuły zostały opublikowane na odpowiedzialność ich autorów. Za treść odpowiadają autorzy poszczególnych tekstów.

ISBN: 978-83-66772-02-1

Opracowanie

Niniejsza książka elektroniczna DVD ma służyć młodym naukowcom. Propagujemy podejmowane działania wśród młodych naukowców, wiedzę, innowacyjne badania oraz rozwój nauki. Nauka musi charakteryzować się ciągłym rozwojem. Dzisiejsi naukowcy korzystają z coraz to nowocześniejszych metod badawczych, prowadzą różnego rodzaju projekty, których efekty w nieodległej przyszłości mają służyć całej społeczności i otaczającemu nas środowisku. Niniejsze opracowanie zawiera zbiór zagadnień prezentujących zainteresowania naukowe młodych adeptów nauki.

Młody naukowiec

Absolwenci studiów drugiego stopnia coraz częściej podejmują decyzję o rozpoczęciu studiów doktoranckich. Decyzja ta często podyktowana jest chęcią pozostania na uczelni w charakterze naukowca i wykładowcy. Niestety po otrzymaniu dyplomu doktora nauk tylko część młodych naukowców pozostanie na uczelni macierzystej. Część młodych doktorów zasili inne uczelnie i jednostki naukowe, a zdecydowana większość rozpocznie kolejny etap swojego życia w instytucjach państwowych i firmach prywatnych. Dlatego też obok realizacji własnych badań naukowych i pisania pracy, doktoranci powinni podjąć wszelkie możliwe działania zmierzające do nawiązania współpracy z firmami prywatnymi, aby realizować dalszą karierę zawodową. Włączanie się doktorantów w różnego rodzaju projekty międzyuczelniane, współpracę w modelu naukowiec--firma, udział we wszelkich konferencjach i szkoleniach o charakterze biznesowo-naukowym zwiększa szanse doktorantów na rozwój naukowy i zawodowy, a przede wszystkim może przynieść upragnioną satysfakcję.

Młodzi naukowcy, którzy pozostali na uczelni wyższej w charakterze często asystenta, adiunkta mają również wiele możliwości nawiązania współpracy ze stale rozwijającym się polskim biznesem. Należy zastanowić się, w jaki sposób przenieść własne dokonania i pomysły naukowe do realizacji w biznesie.

Biznes

Niewątpliwie szansą dla biznesu są innowacje, które niosą ze sobą między innymi młodzi naukowcy. Każdy dobry biznesmen powinien zdać sobie sprawę, że nie ma innowacji bez nowych pomysłów i badań naukowych.

Sami spróbujmy zachęcić właścicieli polskich firm, osoby decyzyjne, menedżerów do nawiązywania współpracy z nami - Młodymi Naukowcami.

Wydawca:
Wydawca: CREATIVETIME, www.creativetime.pl
biuro@creativetime.pl
Skrytka Pocztowa nr 92, 30-093 Kraków 23

Nakład 55 egzemplarzy

Wydanie ISBN

ZAGADNIENIA AKTUALNIE PORUSZANE PRZEZ MŁODYCH NAUKOWCÓW 18

Wydawca: CREATIVETIME

Kraków 2021

BEYOND THE BARRIER OF MISTRUST - AN OVERVIEW OF SELECTED METHODS FOR EXPLAINING PREDICTIONS OF MACHINE LEARNING MODELS

Filip Wójcik

Abstract: This paper reviews some of the existing explanation techniques aimed to clarify machine learning model predictions. It begins with the recap of terminology and unification of some exiting definitions of loosely-defined terms like "interpretability" or "justification". Next, three clarification methods are presented - LIME, SHAP, and ANCHOR, with an intuitive explanation of how they operate and simplified formalization. In the end, a subjective comparison of the methods is presented in terms of theoretical "explainability" postulates. Conducted analysis indicates that ANCHOR is built on top of the easiest to understand mathematical apparatus, while SHAP possesses the most reliable theoretical foundations. The paper concludes with a discussion on the very correctness of applying such solutions instead of bringing focus to transparent models in the first place.

Keywords: machine learning, explainable AI

1. Introduction

With the increasing popularity of machine learning models as decision support systems, the problem of their interpretability and predictive transparency started to rise (Gunning et al., 2019). This inclination has become more visible as deep neural network models started to achieve results beyond humans' capabilities and other machine learning techniques at the cost of their explainability (Holzinger, 2018). The introduction of data protection regulations (like GDPR) was one of the reasons to start the discussion about "XAI" - "Explainable AI" techniques, which might comply with the new regulations (Holzinger, 2018). Areas of social life that require special care and sensitivity (e.g., medicine, justice, high-risk market trading) that benefit from modern AI systems should have access to model-clarification techniques. Some authors even suggest that methods without a simple explanation and transparent structure should not be used in such places, as it poses a risk to privacy and equality and creates a potential for human rights violation (Rudin, 2019).

The next sections present a taxonomy of model-explanation techniques and compare selected approaches in terms of their mechanics.

2. Materials & methods

The following sections present the most important concepts, terminology, and research, related to the subject in question.

Interpretability, explainability and justifications

In the context of machine learning, the terms "interpretability," "explainability," and "justification" are frequently used in exchange. At the same time, no clear definition is provided.

For machine learning, the term "interpretability" can be defined as "an ability to explain or to present in understandable terms to a human" (Doshi-Velez & Kim, 2017). Other authors define it as "the degree to which an observer can understand the cause of a decision" (Miller, 2017). Particular AI models and methods are said to be "interpretable" if "their operations can be understood by a human, either through introspection or through a produced explanation" (Biran & Cotton, 2017). Others formulate "interpretability" as a situation where "a user can correctly and efficiently predict the method's results" (Kim et al., 2016). Some authors (Lipton, 2018) point out that the term itself is fuzzy and poorly defined in scientific terms.

The term "explainability" or "explanation" also is not clearly defined, and (so far) no single, formal definition exists (Molnar, 2019). In philosophical terms of causality, the explanation can be described as follows: "To explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event — explanatory information, I shall call it — tries to convey it to someone else." (Lewis, 1986). Other authors (Miller, 2017) decompose explainability into several factors:

- Cognitive process – an explanation as searching for the description of steps that caused a particular event. In this context, "explanation" is the process of identification of factors and their contribution to the event.
- Product – an explanation as a result of the beforementioned process.
- Social process – an explanation as a knowledge-transfer process, happening between an explainer and explainee.

The explanation is usually placed in time after the model or method has already been trained. Therefore some authors (Lipton, 2018) call them "post hoc" interpretability. In that sense, "explanation" resembles the meaning of the word "justification" - it does not need to clarify the full decision-making process, only its end-result, and why it is considered acceptable or unacceptable (Biran & Cotton, 2017).

In conclusion, during discussions about Explainable AI (XAI), the most relevant set of terms will be "explanation" or "justification", as the complicated models (like neural networks) are not interpretable on their own (Lipton, 2018). Therefore, the ultimate goal is to make such models "post hoc interpretable" using specialized tools or algorithms, providing explanations.

Taxonomy of model explainers

Model explanations can be divided into categories based on how they operate and approach the problem in question.

The first distinction takes into account the point in time at which the explanation is created. Intrinsic explanation methods are built-into the model itself and enforce it to have a more straightforward, understandable structure or to prepare an explanation report during a training procedure (Molnar, 2019). Post-hoc explanations are applied after the model is created, and they do not actively change the model behavior during the training procedure (Lipton, 2018; Molnar, 2019).

The second distinction is based on the explanation tool character. Model-specific approaches are limited to a single model type, while model-agnostic can be applied to any family or kind of machine learning models (Molnar, 2019; Robnik-Šikonja & Bohanec, 2018). For example, interpretation of regression weights in terms of a change in the response variable, when varying specific regressor by a single unit of measure - is definitely a model (regression) specific approach. On the other hand, specialized meta-algorithms described in a "Results" section can explain any family of models, therefore considered "model-agnostic".

The last distinction concentrates on a scope of model explanation. The global explanation is a possibility to understand what general patterns are present in a model (Doshi-Velez & Kim, 2017). Such patterns can be divided even more into:

- Description of a model as a whole at once, so the whole model structure is clear immediately (Lipton, 2018).
- Description of model sub-areas or specific parts (like single trees in a Random Forest or single layers in a neural network) can be used when the model cannot be interpreted at once but needs to be decomposed on a modular level (Molnar, 2019).

On the other hand, local explanations focus on providing explanations for a single data point/entry or entity, considering the models' decisions and the data point surrounding (neighborhood) (Lipton, 2018). Such methods could also be extended to groups of points or entities, providing the areas at which model predictions can be clarified (Molnar, 2019).

Properties of explanations and their quality measures

Explanations should not be too complicated or cryptic, as their goal is to clarify the machine learning model. Therefore, in general, they should possess the following properties (Robnik-Šikonja & Bohanec, 2018):

- Expressive power - understood as being easy to understand and expressive;
- Translucency - the degree to which explanation relies on the model parameters and structure, instead of indirect reasoning via external conditions manipulation or additional (simpler) model training. The example described before - interpretation of linear regression weights possesses a full translucency. The method like "permutation feature importance" (randomly flipping data features to quantify the influence on prediction) (Janitza et al., 2013) possesses none (Lipton, 2018; Molnar, 2019).
- Portability - the range of machine learning model families to which a given explanation method can be applied.
- Algorithmic complexity - is the computational complexity and effort needed to apply a given explanation method.

Quality of explanations can be measured using the following tools (Robnik-Šikonja & Bohanec, 2018):

- Accuracy - the possibility to generalize and port certain explanations to other, unseen yet data points or entities.
- Fidelity - is the degree to which explanations reflect the behavior of the machine learning model in question.
- Consistency - measures how similar are explanations generated for different machine learning model instances, trained on the same problem. The expectation is that explanations for similar models should also be similar.
- Stability - is closely related to consistency but applies to the explanation method itself. The same explanation method applied to the same model multiple times should result in similar conclusions.
- Comprehensibility - defines how well humans can understand presented explanations. This criterion is not easy to quantify and measure, as the understanding is very subjective.
- Certainty - explanation should consider the degree of confidence that the model has for a given prediction.
- Degree of importance - the explanation should also include a description of feature importance.

- Novelty - good explanation should include information if the model can handle unseen data points yet, or if the explanation was based only on training data. This aspect is crucial for judging future performance and the level of generalization.
- Representativeness - how many data points or entities were used to produce an explanation. How representative of the whole training dataset it is?

Some of the criteria mentioned above are hard to quantify or express in terms of numbers. Therefore a large portion of explanation tools and methods are subjective in judgment.

One additional term that is a key concept in the next sections is the so-called "black-box machine learning model". In the literature, the term is attributed to models that are hard to understand by humans, or their internal structure is unreadable without specialized methods (Lipton, 2018). In other words, "in machine learning, black box, describes models that cannot be understood by looking at their parameters" (Molnar, 2019).

3. Results

Sections below describe some of the selected post hoc explanation tools and techniques used in machine learning. They are exemplifications of general concepts described earlier and can serve as a starting point for future study.

LIME

LIME (Local Interpretable Model-agnostic Explanations), as the name suggests, is the model-agnostic, external algorithm capable of providing local explanations (Ribeiro et al., 2016). The explanation procedure is based on random sampling predictions of a black-box model in question and building a new interpretable model on top of such dataset. "An interpretable model" (explainer) is usually selected from the regression algorithms family, as their parameters can directly be related to particular features in a dataset (Ribeiro et al., 2016; Robnik-Šikonja & Bohanec, 2018). This procedure can be formalized as follows:

$$\xi(x) = \arg\min_{g\in\mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad [1]$$

with

$\xi(x)$	the best explanation model for a data point x
$\mathcal{G}$	family of interpretable explanation models (like decision trees, regression, etc.)
$\mathcal{L}$	a measure how "unfaithful" is the explanation model *g* to the original black-box model
f	a measure function
π_x	proximity (neighborhood) function for the point x
$\Omega(g)$	complexity penalty

Source: (Ribeiro et al., 2016)

The simplified LIME training algorithm can be described as follows (Molnar, 2019):

- Train the black-box model
- Select the instance (a data point) of interest for which you want to build explanation
- Perturb the dataset and get the black-box predictions for new points
- Weight these points by their distance (proximity) from x
- Train a weighted, interpretable model on such weighted points (e.g. linear regression)
- Explain the prediction by interpreting such a local model (e.g. "classic" interpretation of regression parameters and their impact)
- Use explanations from point 6. to (indirectly) explain the true black-box model.

LIME is applicable to a wide variety of possible scenarios - from tabular data, natural language processing (when using text-embeddings) to image recognition neural networks, where a local explanation is based on the "neighborhood" of an image part (Ribeiro et al., 2016; Robnik-Šikonja & Bohanec, 2018).

This method also has some serious drawbacks. One of the most serious ones is the instability of explanations. Subsequent calls to the LIME procedure applied to the same data point give varying results (Alvarez-Melis & Jaakkola, 2018). Since it is based on a sampling from a "local neighborhood", it is very susceptible to the "curse of dimensionality", which can be considered a severe drawback in problems like natural language processing with embeddings (Robnik-Šikonja & Bohanec, 2018). Moreover, LIME does not directly address the problem of feature interactions, leaving it to the interpretable explainer, which might not detect any relevant patterns due to its (intended) simplicity (Molnar, 2019; Robnik-Šikonja & Bohanec, 2018).

LIME explanations for tabular data can be depicted in the figure below. The source dataset was "Adult Census" for binary classification using social-status data (yearly earnings >=50K USD or <50K USD). Each row in a figure represents a feature, and bars represent an increase/decrease in its contribution to a given class (in this case: "False") prediction. E.g. marital status=married increases probability of belonging to class ">=50K USD" (green bar) while "capital gain <= 0" strongly favors class "<=50K USD", which makes an intuitive sense.

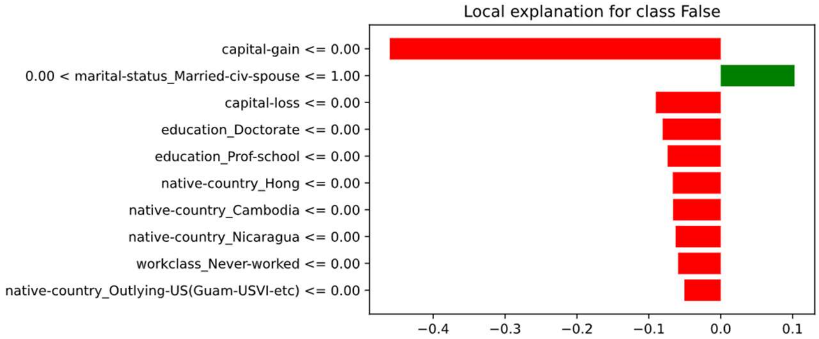


Fig.1: LIME explanations for a tabular dataset, based on "Adult Census" binary classification
Source: own work.

Image explanations based on picture areas, contributing to the target class predictions are presented below.

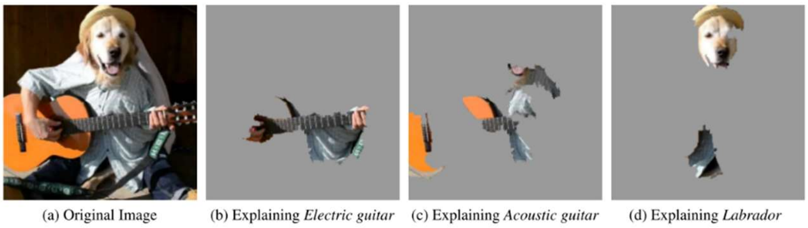

(a) Original Image
(b) Explaining *Electric guitar*
(c) Explaining *Acoustic guitar*
(d) Explaining *Labrador*

Fig.2: Image class prediction using LIME
Source: (Ribeiro et al., 2016).

SHAP

Another approach, similar to LIME is called SHAP (SHapley Additive exPlanations) proposed by Lundberg and Lee (Lundberg & Lee, 2017). Unlike the procedure mentioned above, SHAP utilizes a coalitional game-theoretic approach to guarantee results' theoretical consistency and stability. It gives a unique solution at the cost of very high computational effort.

On a very high level, SHAP utilizes Shapley Values, a weighted average of marginal "player" contributions to the game payouts (Molnar, 2019). In the context of explaining model predictions - each "player" is a feature, attribute, or multiple attributes combination. "Payout" in that circumstance - is their influence on the result. SHAP computes the average feature contribution by checking the prediction value if the feature is used ("active") or not used ("inactive") in a particular prediction ("game") (Štrumbelj & Kononenko, 2014). As features interact, there are multiple combinations, which can be formalized as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_s) \right] \quad [2]$$

with

ϕ_i	Shapley value for i-th feature
$S \subseteq F$	all feature subsets, F is a set of all possible features
$x_S, x_{S\cup\{i\}}$	current data point of interest with i-th feature ignored / enabled
$f_{S\cup\{i\}}(x)$	prediction for a current datapoint from a model with i-th feature enabled
$f_S(x)$	prediction for a current datapoint from a model with i-th feature ignored

Source: (Lundberg & Lee, 2017)

For example, for a simple linear model, there are 2^n possible combinations (Štrumbelj & Kononenko, 2014). Careful analysis of the equation [2] reveals, that the computational effort for computing that equation might fall beyond any computer capabilities. Therefore some approximations (like the assumption of feature independence or integration over samples having particular feature value) were used in practice (Lundberg & Lee, 2017). There are particular SHAP implementations dedicated to individual algorithm families - like TreeSHAP for Decision trees and forests.

The main advantages of SHAP, compared to LIME, are its solid theoretical foundations, uniqueness of the solution, stability, and consistency. On the other hand, the major drawback is a considerable compute time (due to the number of combinations) and complicated mathematical apparatus needed to understand the concept. The former problem can be addressed by using specialized implementation (if it is accessible).

SHAP predictions for tabular data are depicted in the figure below. The source dataset was "Adult Census" for binary classification using social-status data (yearly earnings >=50K USD or <50K USD). The plot presents an effect on a single data point. Red color indicates the influence of the feature towards the "true" class (earnings >= 50K USD).

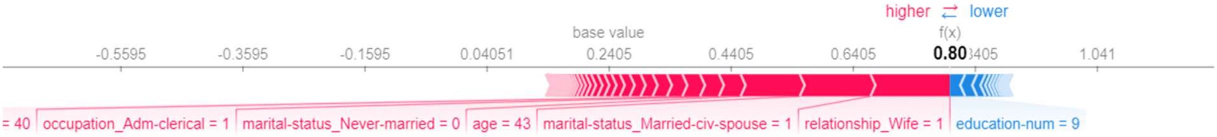


Fig.3: SHAP values for a single datapoint
Source: own work

The figure below shows the summary for all features across the dataset: red color indicates the influence of the feature towards the "true" class (earnings >= 50K USD)

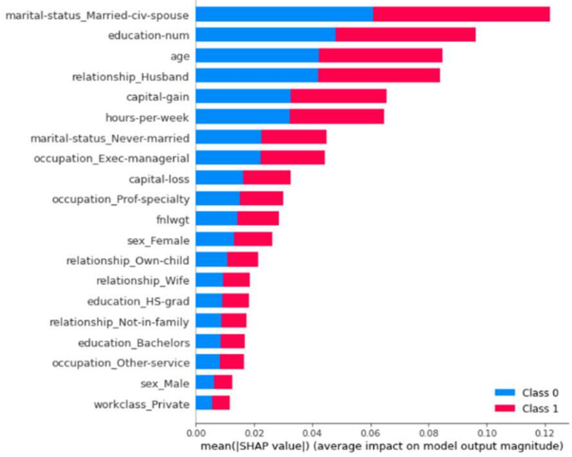


Fig 4: SHAP values for multiple features.
Source: own work

ANCHOR

In this review, the last method is called "Anchor" and was designed by the same scientists who introduced LIME (Ribeiro et al., 2018). ANCHOR's main idea was to provide end-users with an easy-to-understand explanation mechanism without using sophisticated mathematical apparatus.

For any black-box model, ANCHOR can generate local explanations based on if-then-else rules, which are easy to understand (Ribeiro et al., 2018) and are built for a selected data point *x*. Rule antecedent ("if" part) is incrementally enriched with the additional conditions until it covers a sufficient amount of sampled examples. Formally, a rule is called an anchor if it satisfies the following conditions:

$$\mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{I}_{f(x)=f(z)}\right] \geq \tau, A(x) = 1 \qquad [3]$$

with

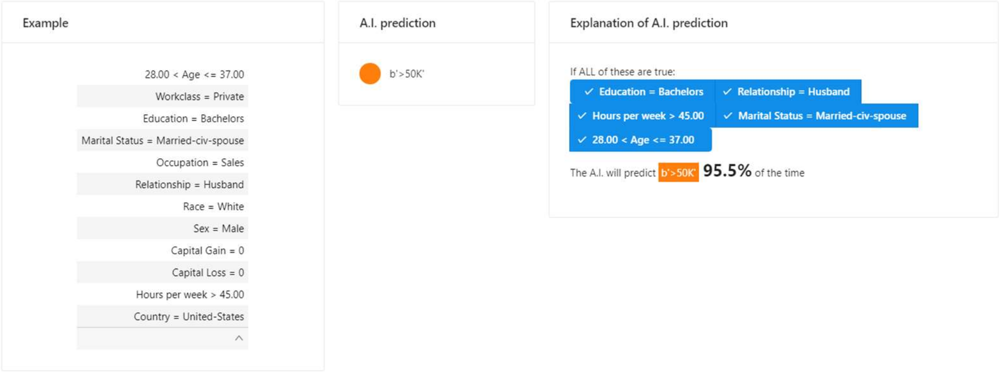
$\mathbb{E}_{D(Z\|A)}$	expected distribution over certain dataset – elements satisfying rule A
A	Anchor rule, a set of predicates (if …) returning 1 if an object x meets criteria
x	data object /record in question
$f(\cdot)$	black-box model prediction on some data point
$\mathcal{D}(\cdot\|A)$	conditional distribution of data points for which rule A applies
$\mathbb{I}_{f(x)=f(z)}$	indicator function returning "true" if rule prediction matches black-box model predictions
τ	threshold value for number of expected points

Source: (Ribeiro et al., 2018)

Simplified anchor rule description can be summarized as follows:
"*Given an instance x to be explained, a rule or an anchor A is to be found, such that it applies to x, while the same class as for x gets predicted for a fraction of at least τ of x's neighbors where the same A is applicable. A rule's precision results from evaluating neighbors or perturbations (following $D(z|A)$) using the provided machine learning model (denoted by the indicator function $\mathbb{I}_{f(x)=f(z)}$).*" (Molnar, 2019).

Such rule-based explanations are "local" and probabilistic, as their so-called "coverage" is defined in terms of the number of covered examples. Rules can be used with different structured data types, e.g., tabular data, text embeddings, or even images, if adequately reformatted. While the generation of all exact rules is infeasible, an optimized beam search procedure is applied to speed up the full process (Molnar, 2019; Ribeiro et al., 2018).

The figure below shows ANCHOR predictions for the same dataset as before.



Example

28.00 < Age <= 37.00
Workclass = Private
Education = Bachelors
Marital Status = Married-civ-spouse
Occupation = Sales
Relationship = Husband
Race = White
Sex = Male
Capital Gain = 0
Capital Loss = 0
Hours per week > 45.00
Country = United-States

A.I. prediction

● b'>50K'

Explanation of A.I. prediction

If ALL of these are true:
✓ Education = Bachelors
✓ Relationship = Husband
✓ Hours per week > 45.00
✓ Marital Status = Married-civ-spouse
✓ 28.00 < Age <= 37.00

The A.I. will predict b'>50K' 95.5% of the time

Fig.5: ANCHOR predictions for sample instance in dataset.
Source: own work.

While it is subjective to judge, which rule visualization is easier to understand, ANCHOR appears to be less complicated because of its rule-based character.

4. Discussion

The explanation tools mentioned above are an exemplification of three main families of justification concepts - an external and local approximation (LIME), external global exact computation with deterministic mathematical component (SHAP), and simplified, external and local, probabilistic rules (ANCHOR). It is hard to quantify how well particular methods satisfy the descriptive conditions described at the beginning of this paper due to the subjective character of understanding. Some authors carried out tests on human subjects to judge the "simulatability" of explanation methods (Hase & Bansal, 2020). Model is "simulatable" if a person can predict its behavior on new, unseen data (Hase & Bansal, 2020; Lipton, 2018). While such a method helps quantify the "novelty" property, it does not cover all the remaining theoretical explainability postulates.

The general properties of the tools described in previous sections can be summarized in the table. Attributes such as "mathematical complexity" are the subjective judgment of the Autor and were based on comparisons between the three methods.

Tab. 1: Explainers comparison

Method	Summary	Computational effort (compared with others)	Solution	Complication of mathematical apparatus
LIME	Local linear models	Moderate	Moderate stability, sampling	Moderate
SHAP	Shapley values	High	High stability, unique theoretical solution	High
ANCHOR	Probabilistic rules	Low	Moderate stability, probabilistic sampling	Low

Source: own work

Correctness, clarity, and other quality measures of explanation tools should be judged in the context of a specific task or problem in question.

The high subjectiveness and complication level of some explanation tools resulted in criticism by some authors. C. Rudin points out that many AI researchers nowadays focus on designing new, accurate black-box models, scoring additional benchmark points, instead of thinking about their explainable alternatives (Rudin, 2019). The main criticism is related to high-stake decisions usage, like criminal justice, healthcare issues, or equality problems. The trend is visible, especially in the USA, where black-box machine learning algorithms are used to predict, e.g., recidivism probability of ex-convicts (Wang et al., 2020). According to that, argumentation, transparency, interpretability, and comprehensibility are of less priority than increasing accuracy metrics (Rudin, 2019). Performance differences between complex, black-box algorithms and their fully transparent counterparts are significant from a statistical point of view, but such differences are much less critical in real-world scenarios (Rudin et al., 2018). Additionally, it is criticized that the methods of explanation are not sufficiently clear and violate most of the theoretical postulates discussed above (Rudin & Radin, 2019; Wang et al., 2020).

5. Conclusions

The discussions about the explainable AI (XAI) concept differ from the other topics in data science, as it is a challenging task to quantify, measure, or even define "interpretability", "explainability" or "justification". In the face of formal regulation like European GDPR, questions about transparency, equality, and the "right to explain and be explained" are more actual than ever (Goodman & Flaxman, 2017). Explanation tools presented in this paper - LIME, SHAP and ANCHOR - are exemplifications of the so-called "external approach" for building post hoc model justifications. They vary in terms of mathematical complexity, visualization techniques, and completeness, and it is up to the end-user in the face of a concrete problem to decide which one to choose. The discussion on the very correctness of applying such solutions is fascinating, as it brings focus to a different topic - should pure accuracy be the ultimate goal of machine learning studies, at the cost of sacrificing transparency and understanding the methods? Further study on easier to understand clarification methods and fully explainable models is required to solve that problem.

6. Literature

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. ArXiv Preprint ArXiv:1806.08049.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. IJCAI-17 Workshop on Explainable AI (XAI), 8.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. ArXiv Preprint ArXiv:1702.08608.

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation.” AI Magazine, 38(3), 50–57.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. Science Robotics.

Hase, P., & Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? ArXiv Preprint ArXiv:2005.01831.

Holzinger, A. (2018). From machine learning to explainable AI. 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), 55–66.

Janitza, S., Strobl, C., & Boulesteix, A.-L. (2013). An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics, 14(1), 1–11.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. Advances in Neural Information Processing Systems, 29, 2280–2288.

Lewis, D. (1986). Causal explanation, in his philosophical papers, Vol. 2. Oxford: Oxford University Press.

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. CoRR - Computing Research Repository, abs/1706.0.

Molnar, C. (2019). Interpretable Machine Learning. Leanpub. https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. AAAI, 18, 1527–1535.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-based explanations of prediction models. In Human and machine learning (pp. 159–175). Springer.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don’t need to? A lesson from an explainable AI competition. Harvard Data Science Review, 1(2).

Rudin, C., Wang, C., & Coker, B. (2018). The age of secrecy and unfairness in recidivism prediction. ArXiv Preprint ArXiv:1811.00731.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647–665.

Wang, C., Han, B., Patel, B., Mohideen, F., & Rudin, C. (2020). In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. ArXiv Preprint ArXiv:2005.04176.

Institution: Wroclaw University of Economics and Business, Faculty of Management Sciences
Research supervisor: dr hab. Iwona Chomiak-Orsa, prof. UE
Correspondence address: filip.wojcik@ue.wroc.pl