THE GRAMMAR OF INTERACTIVE EXPLANATORY MODEL ANALYSIS

A PREPRINT

Hubert Baniecki
Faculty of Mathematics and Information Science
Warsaw University of Technology
hbaniecki@gmail.com
https://orcid.org/0000-0001-6661-5364

Przemyslaw Biecek
Faculty of Mathematics and Information Science
Warsaw University of Technology
przemyslaw.biecek@gmail.com
https://orcid.org/0000-0001-8423-1823

May 4, 2020

ABSTRACT

When analysing a complex system, very often an answer for one question raises new questions. The same law applies to the analysis of Machine Learning (ML) models. One method to explain the model is not enough because different questions and different stakeholders need different approaches. Most of the proposed methods for eXplainable Artificial Intelligence (XAI) focus on a single aspect of model behaviour. However, we cannot sufficiently explain a complex model using a single method that gives only one perspective. Isolated explanations are prone to misunderstanding, which inevitably leads to wrong reasoning.

In this paper, we present the problem of model explainability as an interactive and sequential explanatory analysis of a model (IEMA). We introduce the grammar of such interactive explanations. We show how different XAI methods complement each other and why it is essential to juxtapose them together. We argue that without multi-faceted interactive explanation, there will be no understanding nor trust for models. The proposed process derives from the theoretical, algorithmic side of the model explanation and aims to embrace ideas learned through research in cognitive sciences. Its grammar is implemented in the modelStudio framework that adopts interactivity, automation and customisablity as its main traits. This thoughtful design addresses the needs of multiple diverse stakeholders, not only ML practitioners.

***K*eywords** eXplainable Artificial Intelligence · Interactive Explanations · Black-Box Models · Human-Oriented XAI · Explanatory Model Analysis · Decision-making

1 Introduction

A rapidly increasing number of Machine Learning (ML) applications has demonstrated high efficiency of complex and flexible predictive models, aka Black Boxes. At the same time, there is a growing awareness among users of these models, that we require better tools for exploration and explanation.

There are a lot of technical discoveries in the field of eXplainable Artificial Intelligence (XAI) praised for their mathematical brilliance and software ingenuity [1, 2]. However, in all this rapid development, we forgot about how important is the interface between human and model. There is a huge margin for improvement in the area of human-oriented XAI [3, 4].

To live comfortably, people must trust models predictions to support their everyday lives and not harm them while doing so. Because of some spectacular AI failures even among the most technologically mature companies (see examples related to Google [5], Amazon [6] or Apple [7]) governments and unions step up to provide guidelines and regulations on AI to ensure its safeness, robustness and transparency [8, 9]. The debate on the necessity of XAI is long over. With a right to an explanation comes great responsibility for everyone creating algorithmic decision-making to deliver some form of proof that this decision is fair [10].

Constructing and assessing such evidence becomes a troublesome and demanding task. Surprisingly we have a growing list of end-to-end frameworks for model development such as TensorFlow [11], PyTorch [12], MLlib [13], mlr [14], H2O [15], caret [16] or scikit-learn [17], yet not that many complete and convenient frameworks for model interpretation, explanation and validation. According to [18], the three leading solutions to black-box problems are: evading it and using interpretable algorithms [19], augmenting white-box surrogate models that are understood [20], or using XAI methods for post hoc explanations. Although the first two are precise, the last solution is of particular interest of ours in this paper.

Focusing on overcoming the opacity in ML has led to the development of many model-agnostic explanations such as SHAP [1], LIME [2], Break-Down [21], ALE [22] or PDP [23]. They focus on explaining a specific aspect of a model and often are supported by open-source contributions. There is a great need to condense many of those explanations into all-around frameworks for ML practitioners. Because of that, numerous technical solutions were born that aim to unify the natural and programming language for model exploration, e.g. DALEX [24], iml [25], Skater [26], ELI5 [27], interpretML [28] or AIX360 [29]. They calculate various local and global level model explanations, which help to understand models predictions next to its overall complex behaviour. It is common practice to produce visualisations of these explanations as it is more straightforward to interpret plots than raw numbers. Despite unquestionable usefulness of XAI frameworks, they have a high entry threshold that requires programming proficiency as well as technical knowledge of ML.

Research in cognitive sciences shows that there is a lot to be gained from the interdisciplinary look at XAI [30]. There is a room for improvement in existing solutions, as most of them rarely take into account the human side of the black-box problem [3]. While developing XAI frameworks, we should take into consideration the needs of multiple diverse stakeholders [31, 32, 33], which might require a thoughtfull development of the user interface [34]. It is a different approach than in the case of ML frameworks, where we mostly care about the view of ML practitioners.

As learned in [35], we can extend XAI designs in many ways to embrace the human-oriented, user-centric approach. For us, the key ideas are: (1) Provide contrastive explanations that cross-compare different aspects of a model. (2) Give exploratory information about the data that hides under the model in question. (3) Integrate multiple explanations into a single, more cohesive dashboard. (4) Support the process with useful, additional factors (e.g. explanation uncertainty, feature correlation).

Such a combination can be achieved through the Interactive Explanatory Analysis process introduced in this paper, thus significantly facilitate our understanding of black-box models. It should be pointed out that we mainly focus on predictive black-box models, trained on tabular data, which is a considerable part of nowadays Machine Learning world.

This article has three main contributions:

1. We introduce the grammar of Interactive Explanatory Model Analysis that goes out towards expectations of current challenges in human-oriented XAI.
2. We present its implementation in the modelStudio[1] open-source library.
3. We compare related works from the perspective of automation and interactivity.

Structure of the paper is the following. We overview challenges in providing meaningful insights on black-box models for multiple ML stakeholders at once (Section 2). We explain what we mean by Interactive Explanatory Model Analysis (Section 3) and present the modelStudio framework (Section 4). Then we showcase related work and compare to similar frameworks (Section 5). To conclude, we sketch possible future advancements for this branch of XAI research (Section 6).

2 Challenges in Human-Oriented XAI

Explaining complex predictive models has a high entry threshold, as it may require:

- **Know-how**: We produce explanations using frameworks that involve high programming skills.
- **Know-why**: We need to understand the algorithmic part of the model and heavy math behind explanations to reason properly.
- **Domain knowledge**: We validate explanations against the domain knowledge.
- **Manual exploration**: We need to approach various aspects of a model and data differently, because *all valid models are alike, and each wrong model is wrong in its way*.

[1] https://github.com/ModelOriented/modelStudio

It is possible to enhance the model explanation process to lower the entry threshold and facilitate the exploration of different aspects of a model. In this section, we introduce three main traits that a modern XAI framework should possess to overcome some of the challenges in the interface between a human and a model.

2.1 Interactivity

Interactive dashboards are a popular bussines intelligence tool for data visualisation and analysis due to their ease of use and instant feedback loop. Decision-makers are enabled to work in an agile manner, avoid producing redundant reports and need less know-how to perform demanding tasks. Unfortunately, this is not the case with XAI tools, where most of the current three-dimmensional outputs are mainly targeted at ML practitioners or field-specialists as oppose to nontechnical users [36]. We should focus on developing interactive model explanations that will better suit wider audiences. Such a fourth dimension helps in the interpretation of raw outputs because users can access more information. Additionally, the experience of using interactive tools is far more engaging for users.

Explanations, even in the form of plots, might not be evident and easy to understand. Automaticaly generated captions in the form of additional descriptions are a valuable addition to these visualisations, especially for fresh users without extensive knowledge in the field. Our experience shows that even groups of specialists such as doctors have difficulty analysing more complex charts. The same applies to presented scores and measures, which can be inconsistent or in some cases, misleading. Interactive features like tooltips can add descriptions to the plots. These allow for a more comprehensive range of people to use the already implemented tools.

2.2 Customisability

Interactivity provides an open window for customisation of presented pieces of information. In our means, customisability allows modifying the explanations dynamically. It means that all of the interested parties can freely view and explore model explanations in their way. This trait is essential because human needs may vary over time or be different for different models. With overcoming of this challenge, we reassure that calculated XAI outputs can be adequately and compactly served to multiple diverse consumers [37].

Furthermore, looking at only a few potential plots or measures is not enough to grasp the whole picture. They may very well contradict each other or only together suggest evident model behaviour. Thus to achieve higher quality interpretation, we should compare local level explanations with global level explanations side by side. It broadens the overall model understanding and promotes more profound discoveries.

2.3 Automation

In the model development process [38], a quick feedback loop is desirable. However, endless, manual and laborious model exploration may be a slow and demanding task. Current software gives numerous options, such as model performance measures, feature importance measures, and fairness scores. Selecting and combining these to achieve the necessary result uses many resources.

Moreover, it is tedious to spend much time and effort on producing the explanations, which significantly extends the gap between humans and AI. Checking up on each created model to provide its reasoning might be hard to achieve with available tools. For this process to be successful and productive, we have to develop fast model debugging methods. By fast, we mean easily reproducible in every iteration of the model development process.

While working in an iterable manner, we often reuse our pipelines to explain the model. This task can be fully automated and allow for more active time in interpreting the explanations. Especially in the context of XAI, analysing the results should take most of the time instead of producing them. Another way to automate this process is to calculate various model explanations during model engineering automatically. Adding such improvement to the already existing ML framework lowers the entry threshold by a lot and provides additional information that can be useful.

Even looking apart from the human-oriented side of this topic, ML practitioners can greatly benefit from exploring models to upgrade their quality. Businesses strive for the best possible model performance while trying to spend less time building the infrastructure. It all adds up to higher revenue and a substantial reduction in costs. XAI tools are often used to support the training of ML models. They are useful when dealing with unclear decision-making, so in an obvious way, they can be helpful while improving the accuracy of these decisions. Because of that, we see some of the ML frameworks adopting state-of-the-art model explanations and even automatically calculate them synchronously while training or predicting outcomes. We think that especially the last idea should be further promoted.

3 The Grammar of Interactive Explanatory Model Analysis

Figure 1 shows how the perception of explanability changes with time. For some time the interpretability of the model was not considered important, only the performance on the test set was counted. The next stage was the first generation of explanations focused on individual aspects of the model. The next generation of explanations will focus on analysis to multiple aspects of a model. Necessary requirements for the second generation of explanations are: well defined taxonomy for explanations, and definition of the grammar generating the sequence of explanations.

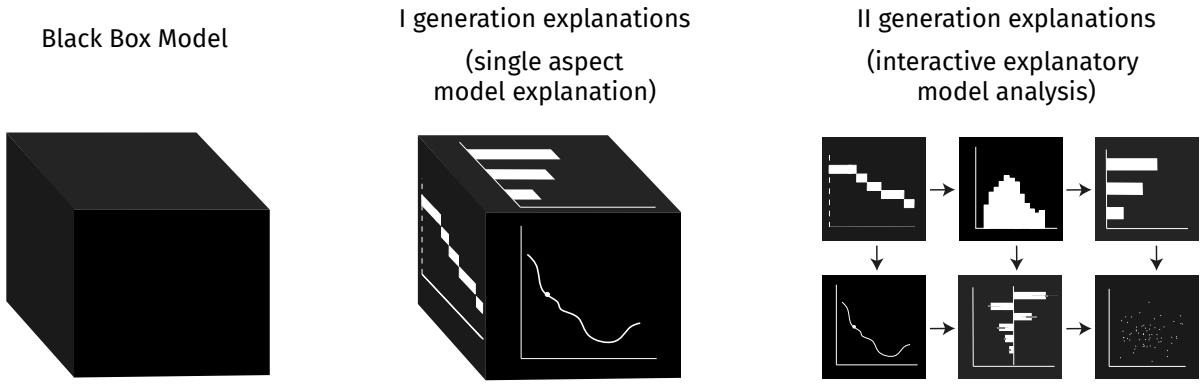

Figure 1: The first generation of model explanations aims at exploring individual aspects of a model behaviour. The second generation of model explanation aims at integration of individual aspects into a vibrant and multi-threaded customisable story about the model that address the needs of different stakeholders.

3.1 Taxonomy of explanations for IEMA

In this subsection, we introduce a new taxonomy of methods for model explanations. Figure 2 shows the two main dimensions of this taxonomy. In the next subsection, on the basis of this taxonomy, we show how different methods can complement each other. The taxonomy is based on two dimensions. The first dimension categorizes the methods according to the question *"What to explain?"*. The second dimension groups the methods according to the question *"How to explain?"*.

The proposed taxonomy distinguishes three groups of explanations in the first dimension. It is consistent with taxonomies introduced in [29, 39, 25].

1. **Data exploration.** These techniques have the longest history (see for example [40]). They focus on the presentation of the distribution of individual variables or relationships between pairs of variables. Often data exploration is conducted to identify outliers or abnormal observations. Data exploration may be interesting to every stakeholder, but most important is for model developers. Understanding data allows to build better models. Examples of such methods are histograms, scatterplots or boxplots.
2. **Global model exploration.** Techniques for model explanations are focused on the behaviour of models on a certain data-set. Unlike data explanations, the main focus here is that we are interested in the behaviour of some particular model. For one dataset we can have many models, which differ in quality and number of variables. Global model explanatory methods can be used by different stakeholders, but most often they are of interest to model validators, which check whether a model behaves as expected. Examples of such methods are Model performance metrics, Variable importance or Partial dependence profiles.
3. **Local model exploration.** These techniques deal with the prediction of the model for a single observation. This type of analysis is useful for detailed model debugging. These explanations can also be presented to end-users of the model to justify the decision proposed by the model. Examples of such methods are Shapley values or Ceteris Paribus profiles.

The second dimension groups the explanation methods based on the nature of the performed analysis. Similarly, we distinguish three groups here.

1. **Analysis of the distribution.** These explanations focus on showing the distribution of certain variables. This makes it easier to understand how typical are certain values.
2. **Analysis of parts.** These explanations focus on the importance of the components of a model. The components are single variables or groups of variables. The model output can be quantified by evaluating the quality of the model or the average response of the model. Examples of such methods are Shapley values or Variable importance.
3. **Analysis of the profile.** These explanations cover the effect of model responses to changes in one or more variables. The result is a profile of a target variable as a function of a selected variable in the input data. Examples of such methods are Partial dependence or Ceteris paribus profiles.

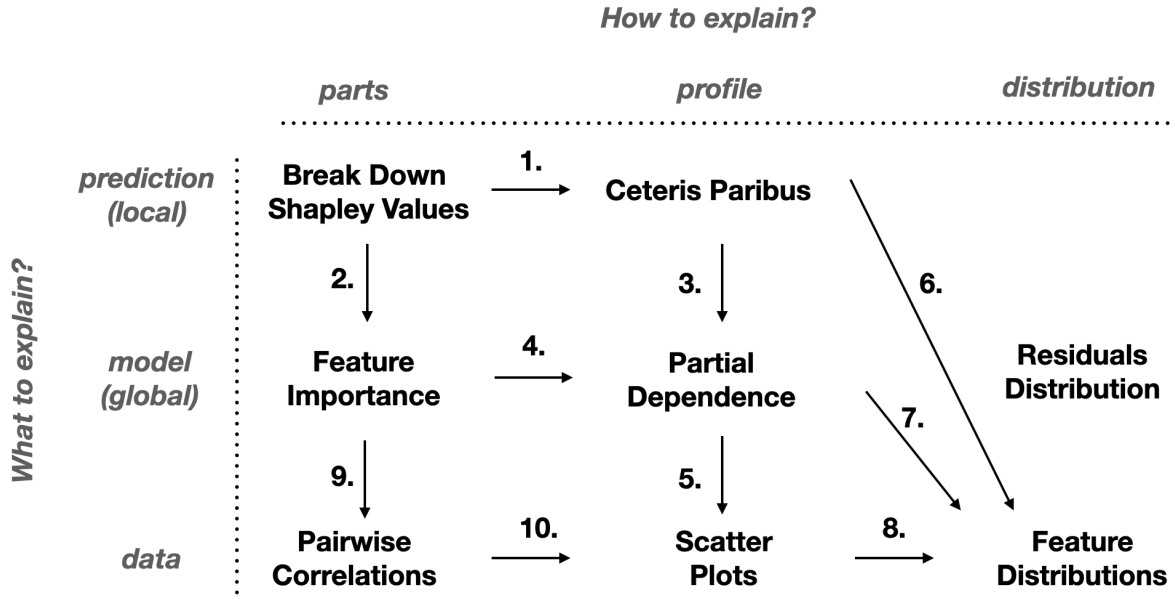

Figure 2: The Grammar of Interactive Model Explanatory Analysis. It shows how the various methods of model exploration enrich each other. Names of popular techniques are listed in cells. Columns and rows span the taxonomy. Edges in this graph indicate which method can be complemented by which.

Figure 2 shows how some well known exploratory techniques fit the proposed taxonomy.

We use the following notation to formalise this taxonomy. Global methods operate on a dataset. Let $\mathcal{D}$ stand for a dataset with n rows and p columns. Here p stands for the number of variables while n stands for the number of observations. Local methods operate on a single observation. Let $x^* \in \mathcal{R}^p$ stand for the observation of interest. Let $f : \mathcal{X} \to \mathcal{R}$ denote for the model of interest, where $\mathcal{X} = \mathcal{R}^p$ is the p-dimensional input space.

When we refer to the analysis of a profile, we are interested in a function that summarises how the model f responds for changes in variable x_i. For local methods such as Ceteris paribus the profile $g(z)$ for variable x_i and observation x^* is defined as

$$g_{x^*}(z) = f(x^*|x_i = z).$$

Global methods such as Partial dependence profile are defined as some aggregation of individual profiles over the whole dataset. For Partial dependence profile $G(z)$ it is an average of Ceteris paribus over all observations x^j

$$G(z) = \sum_{j=1}^{n} g_{x^j}(z).$$

When we refer to the analysis of parts, we are interested in the attribution of some measure to individual variables. For local methods, such as Shapley values, we ask for attributions $h(i)$ for variables x_i that sum up to a model response for data point x^*

$$\sum_{i=1}^{p} h(i) = f(x^*).$$

3.2 Complementary explanations in IEMA

The main results of this paper are based on the observation that each explanation generates further cognitive questions. Model exploration adds up to chains of questions joined with the explanations of different types. This juxtapositioning of different explanations helps us to better understand the behaviour of the model itself.

The explanatory techniques presented in the previous subsection are focused on explaining a single perspective of the model. However, they are not sufficient because every answer raises new questions.

Therefore, when designing a system for explanations, we should also plan possible paths between aspects of a model that complement each other.

In this paper we define interactions with the ML system as a set of possible paths between different aspects of the model. Figure 2 shows a proposed graph of interactions. It creates the grammar of interactive exploration. The edge in the graph means that the selected two aspects of the explanations complete their content. For example Figure 3 shows an example for edge 1, Figure 4 shows an example for edge 6, while Figure 5 shows an example for edge 3.

3.3 Use-case: FIFA 20

We have already introduced the taxonomy of methods for model explanations and the grammar of multi-aspects model explanations. Now, we will present these developments based on the evident data example. There is a regression problem associated with the FIFA 20 dataset [41]. We want to estimate the worth of a player based on the player's characteristics. For this example, a Gradient Boosting Machine model will be explained using the IEMA approach. We use model-agnostic explanations so it could be any other predictive model. Since its structure is irrelevant, we will refer to it as a *back-box* model.

The introduced grammar allows for the construction of the sequence of questions and associated answers. In the case of our model, we will start with a prediction of the worth of one of the most famous footballers, Christiano Ronaldo. The black-box model estimates the value of CR7 at 38M Euro.

Consider the following human-model dialogue:

1. First question: *What factors have the greatest influence on the estimation of the worth of Christiano Ronaldo?* In the taxonomy, this is the local level question about parts. To answer this question, we may present Shapley values or Break down techniques as in Figure 3. The `movement_reactions` and `skill-ball-control` variable increases worth the most, while the `age` is the only variable that decreases Ronaldo's worth.

2. This suggests another question: *What is the relationship between age and the worth of CR7? What would the valuation be if CR7 was younger or older?* This is a local level question about the profile. As the answer, we can present Ceteris paribus technique as in Figure 4. Between the extreme values of the age, the worth differs more than five times.

3. This, in turn, raises the question: *How many players are Ronaldo's age?* In the proposed taxonomy it is a global level question about the distribution. The answer can be the histogram as presented in Figure 4. We see that the vast majority of players in the data are younger than CR7.

4. Another question that may arise is: *Whether such relation between age and worth is typical for other players?* In taxonomy, it is a global level question about the profile. The answer may be a Partial dependence profile, as presented in Figure 5. It is a global pattern that age reduces the worth (with established skills) about five times.

5. However, we know that younger players have lower skills, so another question arises: *What is the relationship between the valuation and age in the original data?* This is the dataset level question about the profile. It is answered by Figure 6.

6. We can also ask which variables are most important when all players are taken into account. This question is answered in Figure 7.

Figures 3-7 show the proces of model exploration. No single explanation will give us as much information about the model as the sequence of various aspects. To keep thoughts flowing, the tool must provide quick feedback-loop between questions. The availability of grammar for IEMA allows for the prior calculation of potential paths between explanations summarised in Figure 8. Such functionality is available in the open-source `modelStudio` tool, which we describe in the next section.

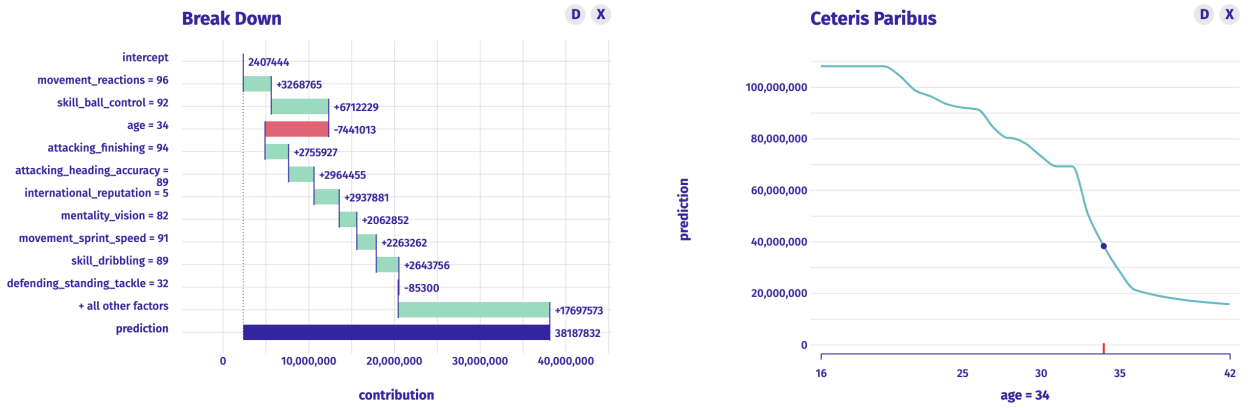

Figure 3: Decomposition of a model prediction (left panel, Break down or Shapley values) shows which variables are most important for a specific instance. It is supplemented by the Ceteris Paribus plot (right panel) which shows the profile response for a specific variable.

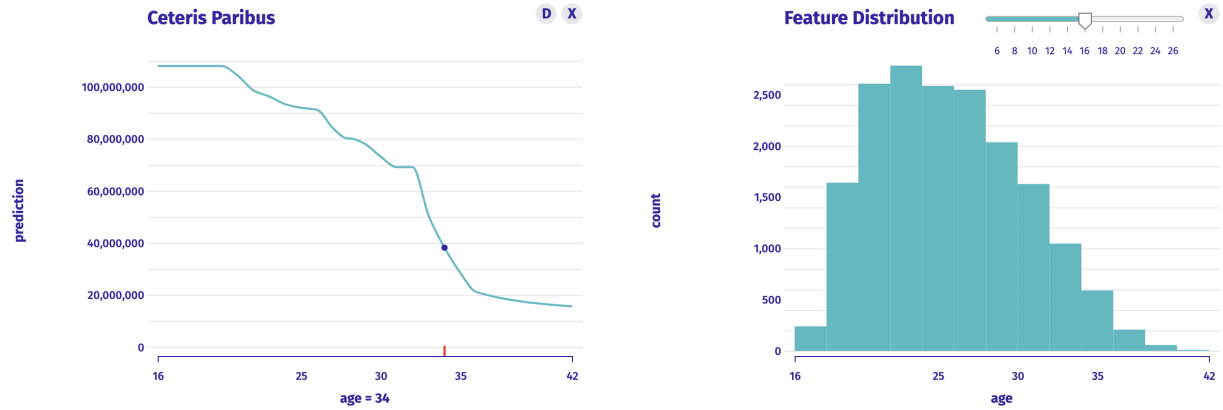

Figure 4: Model response profile for the age variable (left panel, Ceteris Paribus) shows for which values of the model response variable are large or small. It can be supplemented by the histogram (right panel) showing the distribution of values for the age variable.

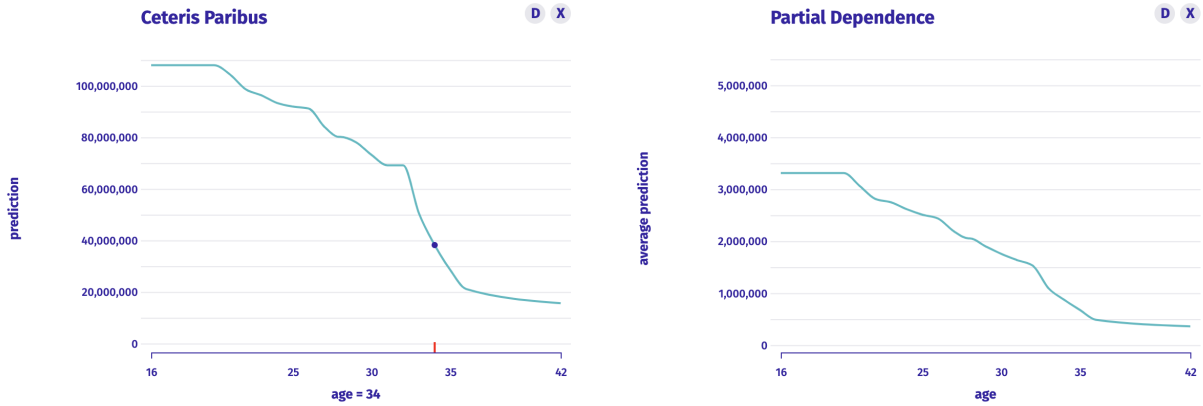

Figure 5: The model response profile for a single instance (left panel, Ceteris Paribus) shows how the model behaves in the neighborhood of that instance. It may be supplemented by an average response profile (right panel, partial dependence).

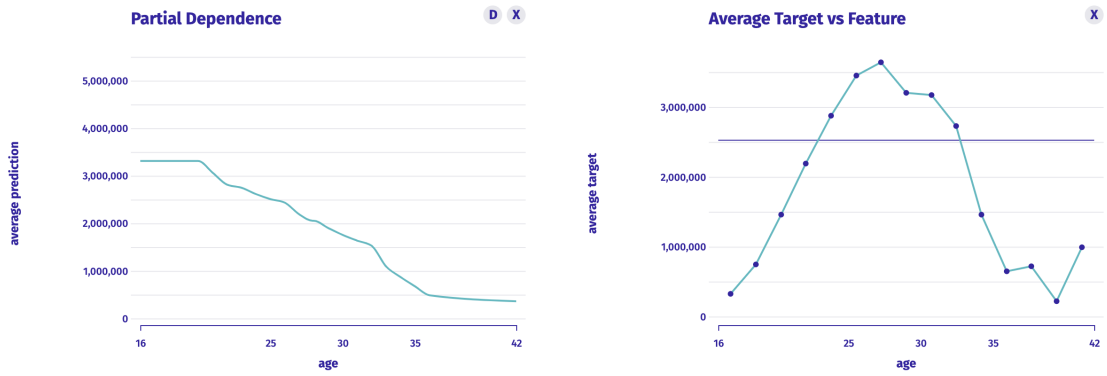

Figure 6: The Partial Dependence profile (left panel) shows the average model behaviour. It can be supplemented by an average value of target variable as a function of selected variable.

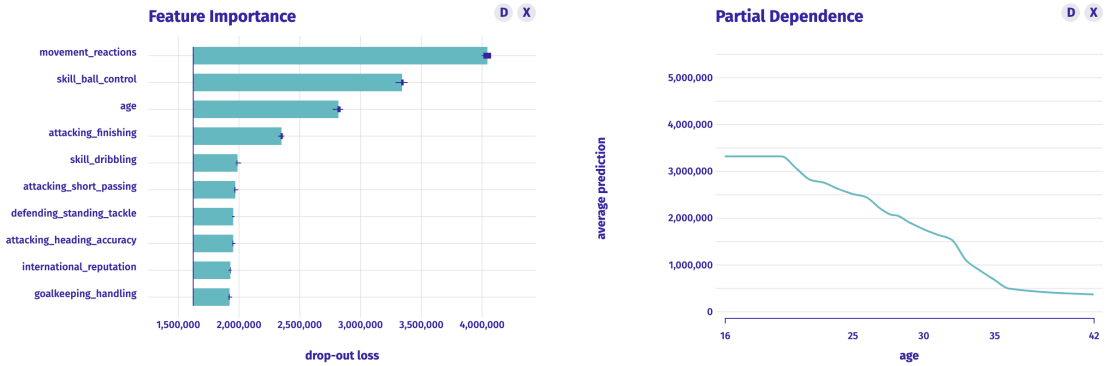

Figure 7: The feature importance plot (left panel) shows which variables influence the model prediction the most. It can be supplemented by an average model response for a selected variable.

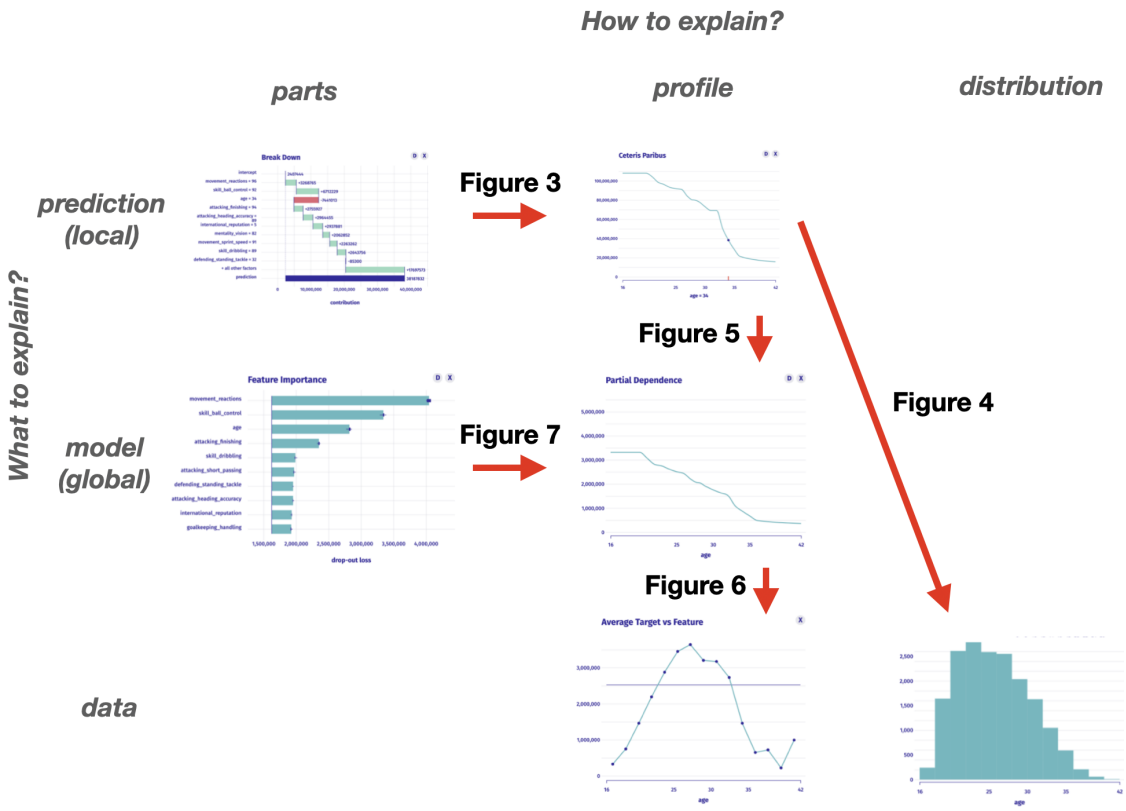

Figure 8: Summary of a single path for interactive model exploration presented in Section 3.3. Different users may choose to explore this graph in different orders.

4 Framework for Interactive Explanatory Model Analysis

In this section, we present a modelStudio framework for Interactive Explanatory Model Analysis. It automatically produces the XAI dashboard consisting of multiple model agnostic explanations. Such a serverless dashboard is easy to save, share and explore by all the interested parties. Interactive features allow for full customisation of the visualisation grid and productive model exploration. Different views presented next to each other broaden the understanding of the path between the model's inputs and outputs, which improves human interpretation of its decisions.

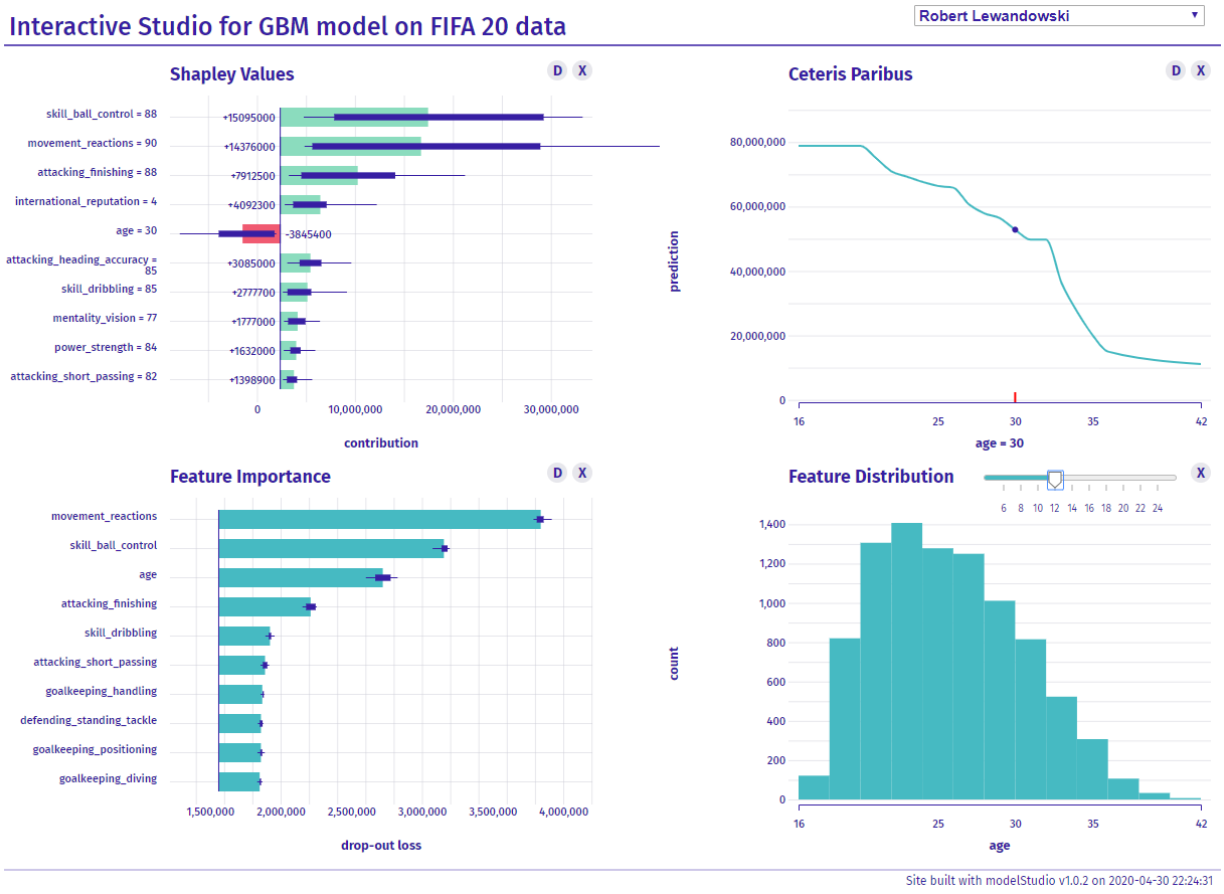

Figure 9: modelStudio automatically produces an HTML file[3] - an interactive and customisable XAI dashboard. Here we present a screenshot of its exemplary layout for the black-box model predicting a player's value on the FIFA 20 data.

The key feature of the output produced with modelStudio is its interactive interface. It is constructed to be user-friendly so that nontechnical users have an easy time navigating through model exploration. There is a possibility to investigate myriad of observations for local model explanations at once, by switching between them freely. The same goes for all of the variables present in the model. No more, there is a need to produce lengthy reports and worry about narrowing down the information.

The juxtaposition of model and prediction level explanations elevates the experience to another level. The whole is greater than the sum of its parts - data distributions accompany parts and profiles. It is a crucial concept to include the necessary data information as a background for a general analysis of models behaviour. Every user can choose a custom grid of panels and change their position at any given time. In Figure 9, we present an example of modelStudio dashboard grid, which consists of Shapley values, Ceteris paribus, Feature importance and Feature distribution plots. One can freely change the data point for prediction parts and profiles with a drop-down box.

Automated natural language descriptions that support visual explanations help to interpret the plots for beginners. They appear after hovering over the plot corner. There is a possibility to add further information, which is relevant for IEMA,

[3] modelStudio output: https://pbiecek.github.io/explainFIFA20/

e.g. true target values of investigated data points, explanations uncertainty via their experimental distributions, and feature correlation[4] mapped interactively on several plots.

This solution puts a vast emphasise on overcoming the challenges discussed in Section 2 and embracing the process presented in Section 3. Overall, working with the produced dashboard is very engaging and effective. modelStudio lowers the entry threshold for all humans that want to understand the black-box models. Due to its automated nature, no sophisticated technical skills are required to produce it. Additionally, it shortens the user-model feedback loop in ML, and creators may efficiently debug models to actively improve their work.

Apart from introduced advantages, the modelStudio output can serve as a supplementary resource for black-box predictive models used in research. The idea of reproducible research is important now more than ever [42, 43]. In the ML domain, there is a debate about adding available data and models as an appendix to research papers. We believe that researchers should also be able to easily support their contributions with model explanations. It would allow others to explore models reasoning and interpret the findings themselves. The modelStudio framework allows for that because its serverless output is simple to produce, save and share.

The same principle stays for ML used in the commercial domain. Decision-making models should have their reasoning put out to the world, and thus make them more transparent for interested parties.

5 Related Work

Here we present work related to this framework. We explicitly omit standard and well-established libraries for model interpretation and explanation as it is a widely documented ground [44]. As discussed in Section 2, they are not entirely going out towards emerging challenges. Although some ideas are discussed in [45], we are looking at tools that recently appeared in this area, especially new developments used in the ML practice. We can divide them into two groups:

1. XAI modules attached to ML frameworks that mostly adopt the automation feature, while also continuously trying to bridge the gap between the humans and AI.
2. Interactive XAI dashboards that focus on treating the model exploration as an extended process and take into account the human side of the black-box problem.

The general incorporation of model explanations into existing ML frameworks is apparent now more than ever. The most popular are the global Feature importance measures. For example, the model-agnostic Feature importance is available in ML libraries [46, 17], while the model-specific feature importance measures often appear next to libraries that focus on a single model [47, 48]. There sparesely are improvements like Partial dependence profiles and Shapley values in such software.

Driverless AI [49] developed by H2O is an all-around state-of-the-art commercial ML platform. It automates feature engineering, model building, visualisation, and interpretability. The last module supports some of the local and global explanations and, most importantly, does not require the user to know how to produce them. While doing a great job at that, it also delivers documentation which describes all of the complex Interpretable ML nuances. The main disadvantages of this framework are its commercial nature and lack of customisation options.

InterpretML [28] developed by Microsoft provides a unified API for model exploration. It can be used to produce explanations for both white-box and black-box models while being compatible with packages like scikit-learn [17]. The ability to create a fully customisable interactive dashboard, that also compares many models at the same time, is a crucial adventage of this tool. Unfortunately, it does not support automation, which, especially for inexperienced people, could be a helpful addition to such a complete package.

TensorFlow [11] developed the TensorBoard [50] dashboard which visualises model behaviour from various angles. It allows tracking models structure, project embeddings to a lower-dimensional space or display audio, image and text data. Furthermore, it promotes adding plugins like the tf-explain [51] library that provides XAI tools tailored for TensorFlow Image Processing models. More related is the What-If Tool [52] developed by Google that allows ML practitioners to explain algorithmic decision-making systems with minimal coding. Using it to join all the metrics and plots into a single, interactive dashboard embraces the grammar of IEMA. What differentiates it from modelStudio is its sophisticated user interface that becomes a barrier for nontechnical users. It also requires a server architecture which might be an inconvenience, as oppose to a serverless modelStudio dashboard.

exBERT [53] is an interactive tool that aims to explain the state-of-the-art Natural Language Processing (NLP) model BERT. It enables users to explore what and how transformers learn to model languages. It is possible to input

[4] Examplary enhancement that will be added soon.

any sentence which will be then parsed into tokens and passed through the model. The attentions and ensuing word embeddings of each encoder are then extracted and displayed for interaction. This example shows a different proposition adapted for the NLP use case but still possesses key traits like automation and interactivity of the dashboard.

Finally, in Table 1 we present a brief comparison of relevant, meaning such as discussed at the start of this Section, XAI frameworks. All of them take a step ahead to provide interactive dashboards with multiple various complementary explanations that allow for a continous model exploration process. Some of these frameworks produce such outputs automatically, which is a high convenience for the user. As stated before, the ultimate XAI framework should be customisable to suit different needs and scenarios. Automation and customisablity make the tool approachable for multiple diverse stakeholders apparent in the XAI domain.

Table 1: Comparison of relevant XAI frameworks. Automated and customisable tools become more approachable for multiple diverse stakeholders, apparent in the XAI domain. Although the What-If Tool partially checks all of the features, it is currently designed for ML practitioners as oppose to nontechnical users.

	modelStudio	Driverless AI	InterpretML	What-If Tool	exBERT
Local Explanation	✓	✓	✓	✓	✓
Global Explanation	✓	✓	✓	✓	
Data Exploration	✓	✓	✓	✓	
Interactive	✓	✓	✓	✓	✓
Automated	✓	✓		✓	✓
Customisable	✓		✓	✓	
Diverse Stakeholders	✓				

6 Conculsions

The topic of eXplainable Artificial Intelligence brings much attention recently. However, the literature is dominated by works either focused on a list of requirements for its better adoption or contributions with a very technical approach to explaining only a single aspect of the model.

In this paper, we propose a third way. First, we argue that explaining a single aspect of the model is incomplete. Second, we propose a taxonomy of methods for explanations, which focuses on the needs of different stakeholders apparent in the lifecycle of Machine Learning models. Third, we describe that Interactive XAI is a process in which explanations are related to a sequence of analysis of complementary model aspects. Fourth, we note that the needs of various users are different; they may also change over time. Therefore, the appropriate interface for unrestricted model exploration must be customisable and accessible to people with proper domain knowledge not necessarily technical ML knowledge.

The introduced grammar of Interactive Explanatory Model Analysis has been designed to allow for effective adoption of a human-oriented approach to XAI. The developed solution `modelStudio` allows for further research on the effectiveness of interactive model exploration.

7 Acknowledgements

We would like to thank Anna Kozak for the design of graphical abstract and Alicja Gosiewska for reviewing this paper. This work was financially supported by the NCN Opus grant 2017/27/B/ST6/01307.

References

[1] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.
URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[2] M. T. Ribeiro, S. Singh, 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, Association for Computing Machinery, 2016, p. 1135–1144.

[3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
URL https://doi.org/10.1016/j.artint.2018.07.007

[4] Z. C. Lipton, The mythos of model interpretability, Queue 16 (3) (2018) 31–57.
URL https://dl.acm.org/doi/abs/10.1145/3236386.3241340

[5] Why google flu is a failure, (Accessed 26 Feb. 2020) (2014).
URL https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/

[6] I. Raji, J. Buolamwini, Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, 2019, pp. 429–435.
URL https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf

[7] Apple's 'sexist' credit card investigated by US regulator, (Accessed 26 Feb. 2020) (2019).
URL https://www.bbc.com/news/business-50365609

[8] ACM U.S. Public Policy Council and ACM Europe Policy Committee, Statement on algorithmic transparency and accountability (2017).
URL https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

[9] The European Commission, White paper on artificial intelligence: a european approach to excellence and trust (02 2020).
URL https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb20_en.pdf

[10] B. Goodman, S. Flaxman, European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation" , AI Magazine 38 (3) (2017) 50–57.
URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/2741

[11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283.
URL https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.
URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-pdf

[13] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al., MLlib: Machine Learning in Apache Spark, Journal of Machine Learning Research 17 (1) (2016) 1–7.
URL http://jmlr.org/papers/v17/15-237.html

[14] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, Z. M. Jones, mlr: Machine Learning in R, Journal of Machine Learning Research 17 (170) (2016) 1–5.
URL http://jmlr.org/papers/v17/15-066.html

[15] H2O.ai, H2O, version 3.28.0.1 (02 2020).
URL http://github.com/h2oai/h2o-3

[16] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software, Articles 28 (5) (2008) 1–26.
URL https://www.jstatsoft.org/v028/i05

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
URL http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[18] P. Hall, N. Gill, N. Schmidt, Proposed Guidelines for the Responsible Use of Explainable Machine Learning (2019). arXiv:1906.03533.

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

[20] A. Gosiewska, P. Biecek, Lifting Interpretability-Performance Trade-off via Automated Feature Engineering (2020). arXiv:2002.04267.

[21] M. Staniak, P. Biecek, Explanations of Model Predictions with live and breakDown Packages, The R Journal 10 (2) (2018) 395–409.

[22] D. W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, CoRR abs/1612.08468 (2019). arXiv:1612.08468.
URL http://arxiv.org/abs/1612.08468

[23] J. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (11 2000). doi:10.1214/aos/1013203451.

[24] P. Biecek, DALEX: Explainers for Complex Predictive Models in R, Journal of Machine Learning Research 19 (84) (2018) 1–5.
URL http://jmlr.org/papers/v19/18-416.html

[25] C. Molnar, G. Casalicchio, B. Bischl, iml: An R package for Interpretable Machine Learning, Journal of Open Source Software 3 (26) (2018) 786.

[26] Skater, version 1.1.2 (02 2020).
URL https://github.com/oracle/Skater

[27] ELI5, version 0.10.1 (02 2020).
URL https://github.com/TeamHG-Memex/eli5

[28] H. Nori, S. Jenkins, P. Koch, R. Caruana, InterpretML: A Unified Framework for Machine Learning Interpretability (2019). arXiv:1909.09223.

[29] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques (2019). arXiv:1909.03012.

[30] M. Westberg, A. Zelvelder, A. Najjar, A Historical Perspective on Cognitive Science and Its Influence on XAI Research, in: Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer International Publishing, Cham, 2019, pp. 205–219.

[31] A. Barredo Arrieta, N. Diaz Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado González, S. Garcia, S. Gil-Lopez, D. Molina, V. R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, Information Fusion (12 2019).

[32] C. Henin, D. Le Métayer, A Multi-layered Approach for Interactive Black-box Explanations, Research Report RR-9331, Inria - Research Centre Grenoble – Rhône-Alpes ; Ecole des Ponts ParisTech (Mar. 2020).
URL https://hal.inria.fr/hal-02498418

[33] K. Sokol, P. Flach, One Explanation Does Not Fit All, KI - Kunstliche Intelligenz (Feb 2020).

[34] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, H. Hussmann, Bringing Transparency Design into Practice, in: 23rd International Conference on Intelligent User Interfaces, IUI '18, Association for Computing Machinery, New York, NY, USA, p. 211–223. doi:10.1145/3172944.3172961.

[35] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing Theory-Driven User-Centric Explainable AI, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019.

[36] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, CoRR abs/1712.00547 (2017). arXiv:1712.00547.

[37] M. Ribera, À. Lapedriza, Can we do better explanations? A proposal of user-centered explainable AI, in: IUI Workshops, 2019.
URL http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

[38] P. Biecek, Model development process (2019). arXiv:1907.04461.

[39] P. Biecek, T. Burzykowski, Explanatory Model Analysis, 2020.
URL https://pbiecek.github.io/ema/

[40] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, 1977.

[41] FIFA 20 dataset at Kaggle, (Accessed 26 Feb. 2020) (2020).
URL https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

[42] G. King, Replication, replication, Political Science and Politics 28 (1995) 444–452.
URL https://j.mp/2oSOXJL

[43] M. Baker, Is there a reproducibility crisis?, Nature 533 (2016) 452–454.
URL https://www.nature.com/news/1.19970

[44] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
URL https://ieeexplore.ieee.org/document/8466590

[45] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Visual Informatics 1 (1) (2017) 48 – 56.
URL http://www.sciencedirect.com/science/article/pii/S2468502X17300086

[46] M. Lang, M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, B. Bischl, mlr3: A modern object-oriented machine learning framework in R, Journal of Open Source Software (dec 2019).
URL https://joss.theoj.org/papers/10.21105/joss.01903

[47] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 3146–3154.
URL http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

[48] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794.
URL https://doi.org/10.1145/2939672.2939785

[49] P. Hall, N. Gill, M. Kurka, W. Phan, Machine Learning Interpretability with H2O Driverless AI, H2O.ai, Inc., 2019.
URL http://docs.h2o.ai

[50] Y. T. Google Inc., TensorBoard, version 2.1.0 (02 2020).
URL https://github.com/tensorflow/tensorboard

[51] tf-explain, version 2.1.0 (02 2020).
URL https://github.com/sicara/tf-explain

[52] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson (Eds.), The What-If Tool: Interactive Probing of Machine Learning Models, 2019.
URL https://arxiv.org/pdf/1907.04135

[53] B. Hoover, H. Strobelt, S. Gehrmann, exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models (2019). arXiv:1910.05276.