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

September 18, 2020

ABSTRACT

When analysing a complex system, very often an answer to one question raises new questions. This also applies to the explanatory analysis of machine learning models. 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. Surprisingly, the majority of methods developed for Explainable Artificial Intelligence (XAI) focus on a single aspect of the model behaviour. In this paper, we show the problem of model explainability as an interactive and sequential analysis of a model. We show how different XAI methods complement each other and why it is essential to juxtapose them together. The proposed process of Interactive Explanatory Model Analysis (IEMA) derives from the theoretical, algorithmic side of the model explanation and aims to embrace ideas developed in cognitive sciences. Its grammar is implemented in the modelStudio framework that adopts interactivity, customisability and automation as its main traits.

Keywords Explainable Artificial Intelligence · Interactive Explanations · Black-Box Models · Human-Oriented XAI · Explanatory Model Analysis · Responsible Artificial Intelligence

1 Introduction

1.1 Background

A rapidly increasing number of machine learning 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 [1–5]. However, in all this rapid development, we forgot about how important is the interface between human and model. Interactive interpreters available in tools such as R or Python significantly facilitate data analysis process. Another breakthrough was notebooks that speed up the feedback loop. Still, there is a huge margin for improvement in the area of human-oriented XAI [6, 7].

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 [8], Amazon [9] or Apple [10]), governments and unions step up to provide guidelines and regulations on AI to ensure its safeness, robustness and transparency [11, 12]. 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 [13].

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 [14], yet not that many complete and convenient frameworks for model interpretation, explanation and validation. According to [15], the three main approaches to black-box model

exploration are: evading it and using interpretable by design algorithms [16], applying discrimination testing techniques [17], or using post-hoc explanation methods. Although the first two are precise, the last solution is of particular interest of ours in this article. We will limit ourselves here to models for tabular data, but the presented approach can also be generalized to other types of data.

Focusing on overcoming the opacity in machine learning has led to the development of many model-agnostic explanations [1, 2, 18–20]. There is a great need to condense many of those explanations into comprehensive frameworks for machine learning practitioners. Because of that, numerous technical solutions were born that aim to unify the natural and programming language for model exploration [3–5]. 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 machine learning.

Research in cognitive sciences shows that there is a lot to be gained from the interdisciplinary look at XAI [21, 22]. 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 [7]. While developing XAI frameworks, we should take into consideration the needs of multiple diverse stakeholders [23–25], which might require a thoughtful development of the user interface [26]. It is a different approach than in the case of machine learning frameworks, where we mostly care about the view of machine learning practitioners.

1.2 Objectives

As learned in [27], 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). In this paper, we introduce grammar of the Interactive Explanatory Analysis, thus significantly facilitate our understanding of black-box models through a sequence of single aspect model explanations.

Structure of the paper is the following. We overview challenges in providing meaningful insights on black-box models for multiple machine learning stakeholders at once (Section 2). We present basics of the grammar of Interactive Explanatory Model Analysis and its implementation in the `modelStudio` framework (Section 4). Finally, we refer to the principles of responsible machine learning and discuss the potential use of our contribution as a defense from adversarial attacks on model explanations (Section 5).

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 which 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*.

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 popular in business intelligence tools 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-dimensional outputs are mainly targeted at machine learning practitioners or field-specialists as oppose to nontechnical users [28]. As an alternative, we could 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.

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 [29]. 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.

2.3 Automation

In the model development process [30], a quick feedback loop is desirable. However, endless, manual and laborious model exploration may be a slow and demanding task. For this process to be successful and productive, we have developed 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.

2.4 Conclusion

Automation and customisablity make the framework approachable for diverse stakeholders apparent in the XAI domain. Interactivity allows for a continuous model exploration process. Standard and well-established libraries for model interpretation and explanation documented in [31] are not entirely going out towards emerging challenges. Although some ideas are discussed in [32], we are relating to tools that recently appeared in this area, especially new developments used in the machine learning practice. They are mostly dashboard-like XAI frameworks that aim to implement the introduced traits [33–40]. We provide a comparison of such tools in Appendix A.

3 The Grammar of Interactive Explanatory Model Analysis

Figure 1 shows how the perception of explainability changes with time. For some time the interpretability of the model was not considered important, the main focus was put to model performance. The next step was the first generation of explanations focused on individual aspects of the model, like the effects of particular variables. The next generation will focus on the analysis of multiple aspects of a model. Requirements for the second generation involve a well-defined taxonomy of model explanations, and definition of the grammar generating their sequences.

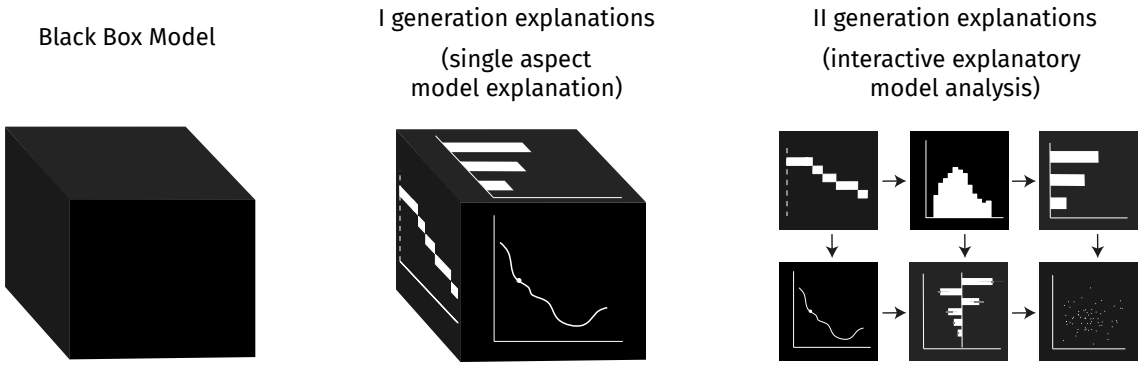

Figure 1: The first generation of model explanations aims at exploring individual aspects of model behaviour. The second generation of model explanation aims at the integration of individual aspects into a vibrant and multi-threaded customisable story about the model that addresses the needs of different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

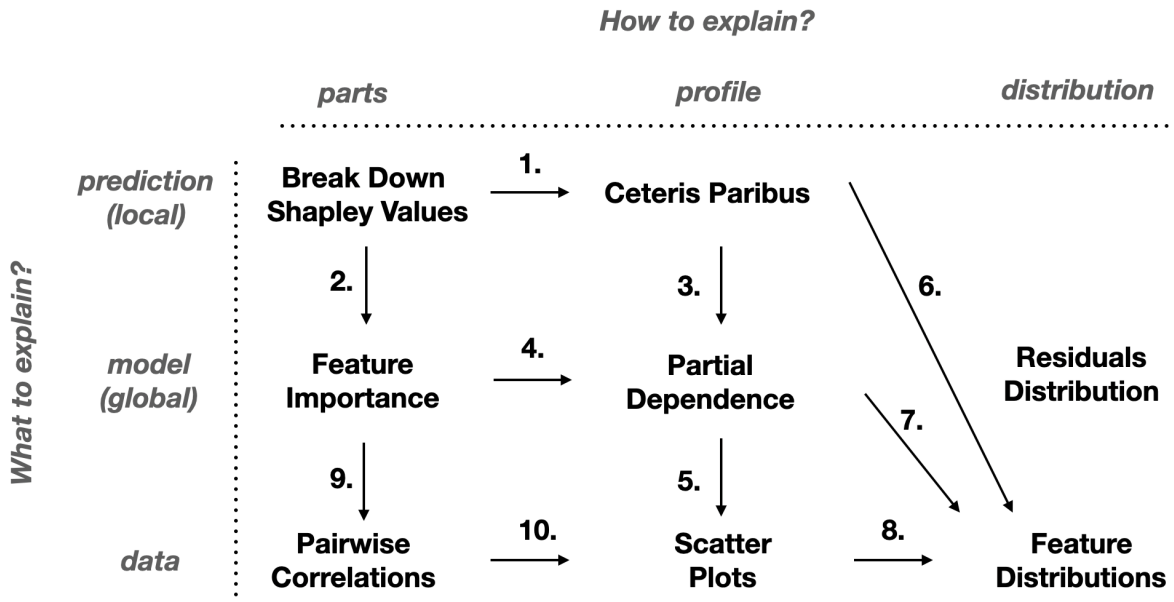

Figure 2: The Grammar of Interactive Model Explanatory Analysis. It shows how the various methods for model explanation enrich each other. Names of popular techniques are listed in cells. Columns and rows span the taxonomy. Edges in this graph indicate which method can be complemented by which.

3.1 Taxonomy of explanations for IEMA

In this subsection, we introduce a new taxonomy of methods for model explanation. 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 *„What to explain?"* question. It is consistent with taxonomies introduced in [4, 5, 41].

1. **Data exploration.** These techniques have the longest history (see for example [42]). 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 them to build better models.
2. **Global model explanation.** Techniques for model explanations are focused on the behaviour of models on a certain dataset. 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, i.e. in the number of variables. Different stakeholders can use global methods, but most often they are of interest to model validators, which check whether a model behaves as expected. Examples of such methods are Performance metrics, Variable importance or Partial dependence profiles.
3. **Local model explanation.** 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. The results make 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 shows how some well known explanatory 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 overall 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. 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 machine learning 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 explanation and the grammar of multi-aspect model explanation. Now, we will present these developments based on the evident data example. There is a regression problem associated with the FIFA 20 dataset [43]. We want to estimate the worth of a player based on his 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.

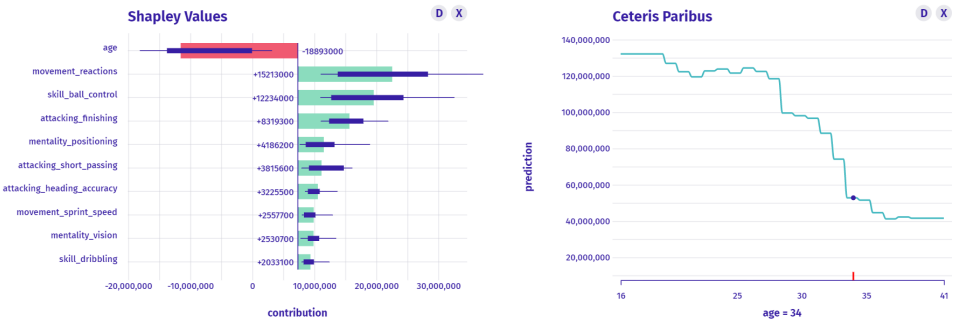

Figure 3: Decomposition of a model prediction (left panel, Shapley values or Break down) shows which variables are most important for a specific instance. It is supplemented by the Ceteris Paribus plot (right panel) which shows the profile response for a specific variable.

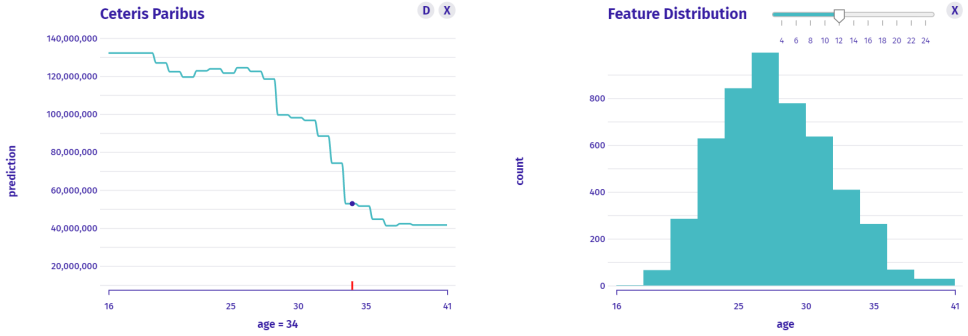

Figure 4: Model response profile for the age variable (left panel, Ceteris Paribus) shows for which values of the model response variable are large or small. It can be supplemented by the histogram (right panel) showing the distribution of values for the age variable.

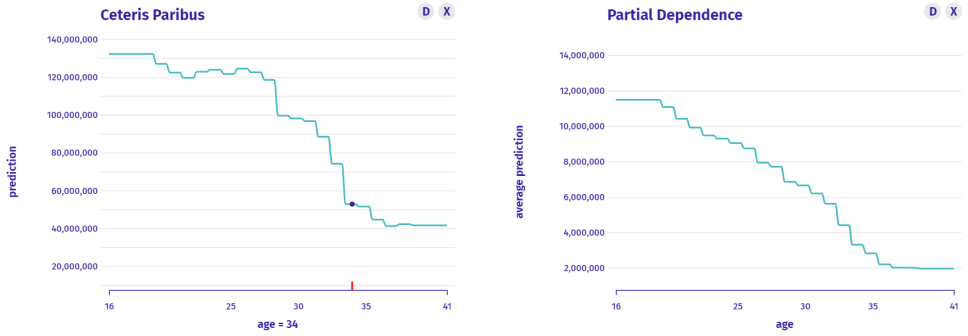

Figure 5: The model response profile for a single instance (left panel, Ceteris Paribus) shows how the model behaves in the neighbourhood of that instance. It may be supplemented by an average response profile (right panel, partial dependence).

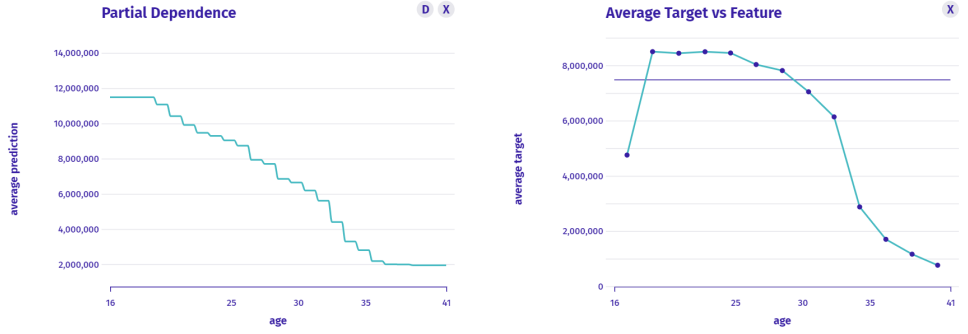

Figure 6: The Partial Dependence profile (left panel) shows the average model behaviour. It can be supplemented by an average value of target variable as a function of selected variable.

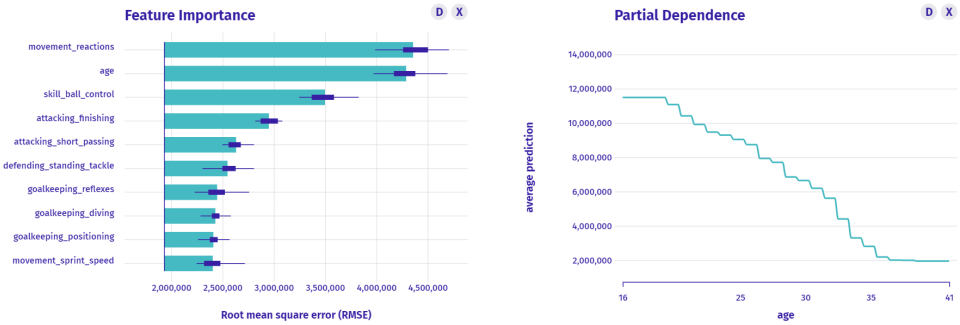

Figure 7: The feature importance plot (left panel) shows which variables influence the model prediction the most. It can be supplemented by an average model response for a selected variable.

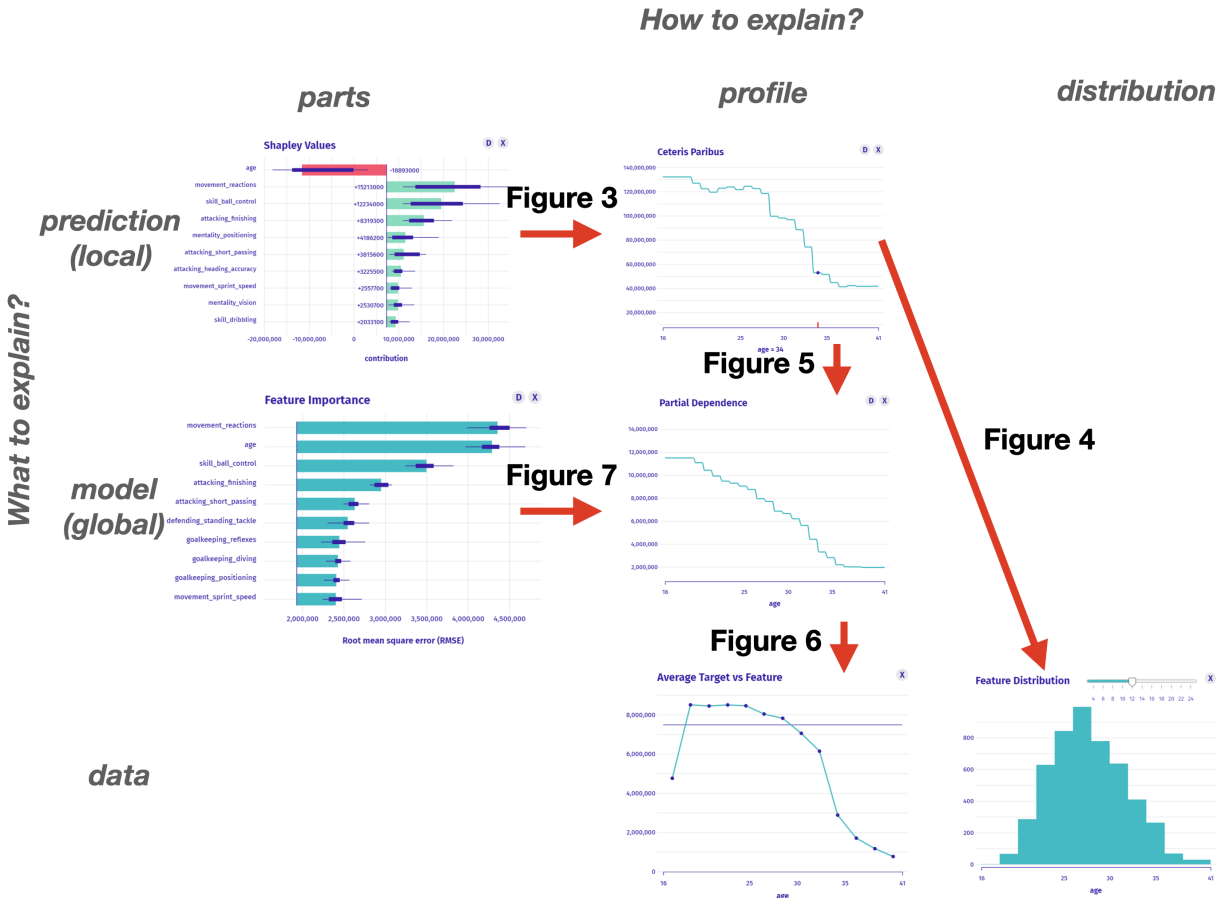

Figure 8: Summary of a single path for interactive model exploration presented in Section 3.3. Different users may choose to explore this graph in different orders.

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. The grammar of IEMA allows for the prior calculation of potential paths between explanations summarised in Figure 8. To keep the thoughts flowing, the desired tool must provide interactive features, customisability and ensure a quick feedback-loop between questions. These functionalities are available in the open-source `modelStudio` framework [33]. Its dashboard output for this use-case is available for everyone to explore on their own at `https://pbiecek.github.io/explainFIFA20/`.

4 Framework for Interactive Explanatory Model Analysis

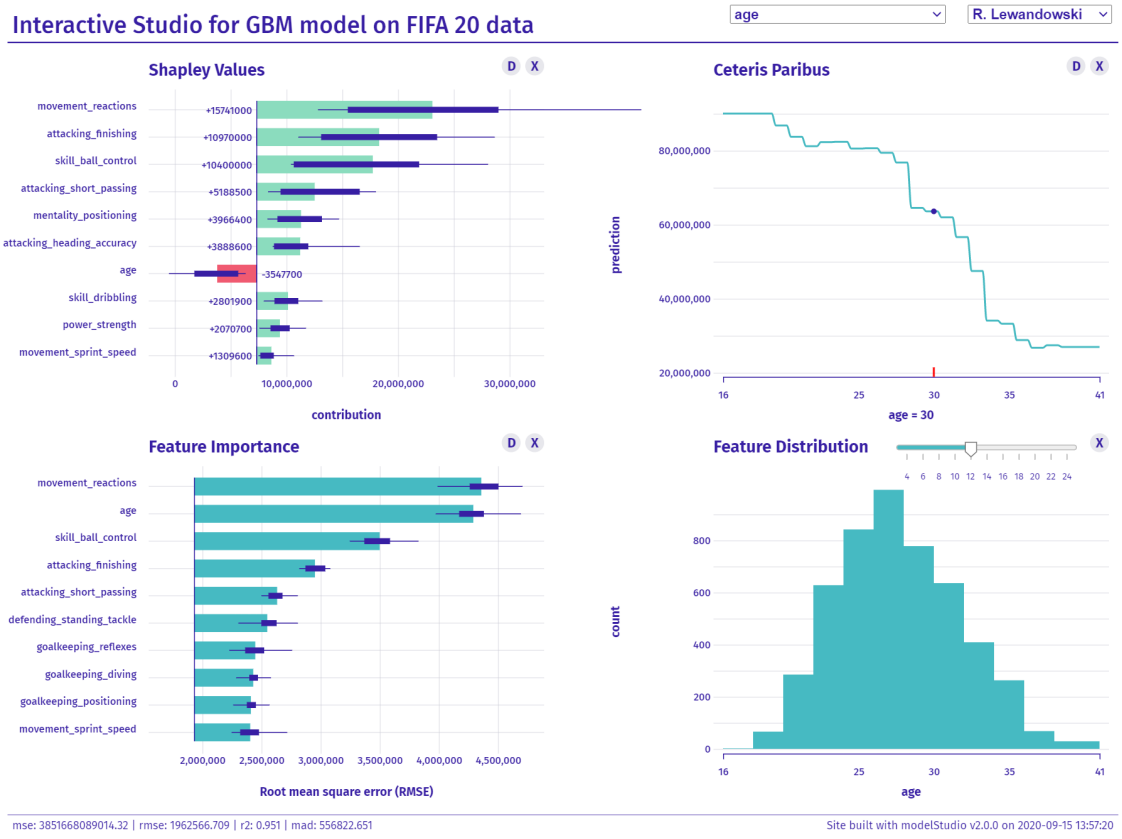

Figure 9: `modelStudio` automatically produces an HTML file - an interactive and customisable dashboard with model explanations and data exploration visualisations. Here we present a screenshot of its exemplary layout for the black-box model predicting a player's value on the FIFA 20 data. See `https://pbiecek.github.io/explainFIFA20/`

The `modelStudio` framework [33] allows performing Interactive Explanatory Model Analysis. It automatically computes various (instance and dataset level) model explanations and produces a customisable dashboard[1], which consists of multiple panels for plots with their short descriptions. These are model agnostic explanations and data exploration visualisations. 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 examination. 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.

The key feature of the output produced with `modelStudio` is its interface. It is constructed to be user-friendly so that non-technical users have an easy time navigating through the process. 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. Any user can choose a custom grid of panels and change their position at any given time.

This solution puts a vast emphasise on overcoming the challenges discussed in Section 2 and implementing 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 the machine learning development stage, and creators may efficiently debug models to actively improve their work. We provide a comparison of similar tools in Appendix A.

5 Discussion

In this section, we discuss further how our contribution corresponds to the recent relevant research topics.

5.1 Responsible machine learning

Responsibility is being brought up as a critical factor in the machine learning domain [15, 23]. An interesting proposition concerning of model transparency, fairness and security is the Model Cards framework introduced in [44]. It aims at providing complete documentation of the model in the form of a short report. There are various information, e.g. textual descriptions, performance benchmarks, model explanations, and valid context. Apart from introduced advantages of the `modelStudio` framework, its output can serve as a supplementary resource for black-box predictive models generated after model development.

The idea of reproducible research is important now more than ever [45, 46]. In the machine learning 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 (especially reviewers) 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 machine learning used in the commercial domain. Decision-making models could have their reasoning put out to the world, and thus make them more transparent for interested parties.

5.2 Adversarial attacks on model explanations

There are various adversarial attacks on machine learning models; hence, ways of defending, e.g. by using XAI techniques. Nowadays, more and more specific attacks on model explanations come to light [47, 48], also in the Computer Vision domain. In [49], authors showcase that a slight manipulation of the input data can result in artificially made model explanations. As a defense to such adversary, [50] presents how a simple aggregation of multiple explanation methods makes the model robust against manipulation.

The idea of *aggregation* of multiple explanation methods is at the core of the IEMA process and the `modelStudio` framework. Not in the technical side, but the integration of individual model aspects into a multi-faceted view of a model. The *juxtaposition* of these aspects gives far more complementary information; thus, XAI methods become robust against deception. We believe, that our contribution has a high potential to be a solid defense against adversarial attacks on isolated model explanations, especially where manipulation of the data is involved, because of the addition of data exploration visualisations.

[1] `modelStudio` dashboard for the FIFA 20 use-case: `https://pbiecek.github.io/explainFIFA20/`
code: `https://github.com/pbiecek/explainFIFA20`

5.3 Conclusions

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 model explanation.

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 model explanation is an interactive process in which we analyse a sequence of complementary model aspects. Therefore, the appropriate interface for unrestricted model exploration must adopt interactivity, customisation, and automation to lower the entry threshold.

The introduced grammar of Interactive Explanatory Model Analysis has been designed to allow for effective adoption of a human-oriented approach to XAI. Its practical implementation is available through the `modelStudio` framework. Such a complete solution could also be utilised as a supplementary resource for black-box predictive models, and as a defense from attacks on model explanations. In the future, we would like to aggregate the data from the user-centric experiments, using the extensive telemetry possibilities of this tool, to look at how different stakeholders analyse the information.

6 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 (Poland) Opus grant 2017/27/B/ST6/01307*.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, page 1135–1144, 2016. URL `https://arxiv.org/pdf/1602.04938.pdf`.

[2] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017. URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

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

[4] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software*, 3(26):786, 2018. URL `https://doi.org/10.21105/joss.00786`.

[5] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. 2019. URL `https://arxiv.org/pdf/1909.03012.pdf`.

[6] Zachary C. Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):31–57, June 2018. URL `https://dl.acm.org/doi/abs/10.1145/3236386.3241340`.

[7] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38, 2019. URL `https://doi.org/10.1016/j.artint.2018.07.007`.

[8] Forbes. Why Google Flu Is A Failure, 2014. URL `https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/`. (Accessed 26 Feb. 2020).

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

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

[11] 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`.

[12] 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-feb2020_en.pdf`.

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

[14] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluch. Machine Learning and Deep Learning Frameworks and Libraries for Large-Scale Data Mining: A Survey. *Artif. Intell. Rev.*, 52(1):77–124, 2019. URL https://doi.org/10.1007/s10462-018-09679-z.

[15] Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information*, 11(3):137, 2020. URL https://www.mdpi.com/2078-2489/11/3/137/htm.

[16] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1:206–215, 05 2019. URL https://doi.org/10.1038/s42256-019-0048-x.

[17] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, 2015. URL https://doi.org/10.1145/2783258.2783311.

[18] Mateusz Staniak and Przemysław Biecek. Explanations of Model Predictions with live and breakDown Packages. *The R Journal*, 10(2):395–409, 2018. URL https://journal.r-project.org/archive/2018/RJ-2018-072/RJ-2018-072.pdf.

[19] Daniel W. Apley and Jingyu Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *CoRR*, abs/1612.08468, 2019. URL http://arxiv.org/abs/1612.08468.

[20] Jerome Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29, 11 2000. URL https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451.

[21] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects. 2018. URL https://arxiv.org/ftp/arxiv/papers/1812/1812.04608.pdf.

[22] Marcus Westberg, Amber Zelvelder, and Amro Najjar. A Historical Perspective on Cognitive Science and Its Influence on XAI Research. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 205–219, 2019. URL http://doi.org/10.1007/978-3-030-30391-4_12.

[23] Alejandro Barredo Arrieta, Natalia Diaz Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado González, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 2019. URL https://doi.org/10.1016/j.inffus.2019.12.012.

[24] Clement Henin and Daniel 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, March 2020. URL https://hal.inria.fr/hal-02498418.

[25] Kacper Sokol and Peter Flach. One Explanation Does Not Fit All. *KI - Künstliche Intelligenz*, 34(2):235–250, Feb 2020. URL http://dx.doi.org/10.1007/s13218-020-00637-y.

[26] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*, page 211–223, 2018. URL http://doi.org/10.1145/3172944.3172961.

[27] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. URL http://doi.org/10.1145/3290605.3300831.

[28] Tim Miller, Piers Howe, and Liz 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. URL https://arxiv.org/pdf/1712.00547.pdf.

[29] Mireia Ribera and Àgata 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.

[30] Przemyslaw Biecek. Model Development Process. 2019. URL https://arxiv.org/pdf/1907.04461.pdf.

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

[32] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48 – 56, 2017. ISSN 2468-502X. URL http://www.sciencedirect.com/science/article/pii/S2468502X17300086.

[33] Hubert Baniecki and Przemyslaw Biecek. modelStudio: Interactive studio with explanations for ML predictive models. *Journal of Open Source Software*, 4(43):1798, Nov 2019. URL https://doi.org/10.21105/joss.01798.

[34] Patrick Hall, Navdeep Gill, Megan Kurka, and Wen Phan. *Machine Learning Interpretability with H2O Driverless AI*. H2O.ai, Inc., October 2019. URL http://docs.h2o.ai.

[35] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability. 2019. URL https://arxiv.org/pdf/1909.09223.pdf.

[36] Yuan Tang Google Inc. *TensorBoard*, 02 2020. URL https://github.com/tensorflow/tensorboard. v2.1.0.

[37] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. 2019. URL https://arxiv.org/pdf/1907.04135.

[38] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models. 2019. URL https://arxiv.org/abs/1910.05276.

[39] *explainX*, 08 2020. URL https://github.com/explainX/explainx. v2.3.6.

[40] *ArenaR*, 08 2020. URL https://github.com/ModelOriented/ArenaR. v0.2.1.

[41] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. 2020. URL https://pbiecek.github.io/ema/.

[42] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[43] FIFA20. FIFA 20 dataset at Kaggle. URL https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset. (Accessed 26 Feb. 2020).

[44] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 2019. URL http://dx.doi.org/10.1145/3287560.3287596.

[45] Gary King. Replication, Replication. *Political Science and Politics*, 28:444–452, 09 1995. URL https://j.mp/2oSOXJL.

[46] Monya Baker. Is there a reproducibility crisis? *Nature*, 533:452–454, 05 2016. URL https://www.nature.com/news/1.19970.

[47] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 161–170, 09-15 Jun 2019. URL http://proceedings.mlr.press/v97/aivodji19a.html.

[48] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, 2020. URL https://doi.org/10.1145/3375627.3375830.

[49] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems 32*, pages 13589–13600. 2019. URL http://papers.nips.cc/paper/9511-explanations-can-be-manipulated-and-geometry-is-to-blame.pdf.

[50] Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations. 2020. URL https://arxiv.org/abs/2007.06381. 2020 Workshop on Human Interpretability in Machine Learning (WHI).

[51] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

[52] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, pages 3146–3154. 2017. URL http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

[53] Brandon M. Greenwell. pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1): 421–436, 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-016/index.html`.

[54] *tf-explain*, 02 2020. URL `https://github.com/sicara/tf-explain`. v2.1.0.

A Comparison of dashboard-like XAI frameworks

Here we present work related to the `modelStudio` framework [33] presented in Section 4. We explicitly omit standard and well-established libraries for model interpretation and explanation as it is a widely documented ground [31]. As discussed in Section 2, they are not entirely going out towards emerging challenges. Although some ideas are discussed in [32], we are looking at tools that recently appeared in this area, especially new developments used in the machine learning practice. We can divide them into two groups:

1. XAI modules attached to the machine learning frameworks that mostly adopt the automation feature, while also continuously trying to bridge the gap between the humans and AI.
2. Interactive dashboard-like tools 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 machine learning workflow is rapidly increasing. The most popular are the global Feature importance measures [51]. For example, the model-agnostic Feature importance is available in comprehensive machine learning frameworks [14], while the model-specific Feature importance measures often appear next to libraries that focus on a single model [52]. There more and more are improvements like Partial dependence profiles [20, 53] and Shapley values [2] in such software.

`Driverless AI` [34] developed by H2O is a comprehensive state-of-the-art commercial machine learning 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 machine learning nuances. The main disadvantages of this framework are its commercial nature and lack of customisation options.

`InterpretML` [35] 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. The ability to create a fully customisable interactive dashboard, that also compares many models at the same time, is a crucial advantage of this tool. Unfortunately, it does not support automation, which, especially for inexperienced people, could be a helpful addition to such a complete package.

`TensorBoard` [36] developed by TensorFlow is a 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` [54] library that provides XAI tools tailored for TensorFlow Image Processing models. More related is the `What-If Tool` [37] developed by Google that allows machine learning 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 non-technical users. It also requires a server architecture which might be an inconvenience, as oppose to a serverless modelStudio dashboard.

`exBERT` [38] 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 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, the newest additions to the list are `explainX` (Python) [39] and `ArenaR` (R) [40] packages. 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 continuous 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 and interactive 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 machine learning practitioners as oppose to non-technical users.

	local explanation	gloval explanation	data exploration	interactive	automated	customisable	diverse stakeholders
modelStudio [33]	✓	✓	✓	✓	✓	✓	✓
Driverless AI [34]	✓	✓	✓	✓	✓		
InterpretML [35]	✓	✓	✓	✓		✓	
Tensorboard [36, 54]	✓		✓	✓	✓		
What-If Tool [37]	✓	✓	✓	✓	✓	✓	
exBERT [38]	✓			✓	✓		
explainX [39]	✓	✓	✓	✓	✓		
ArenaR [40]	✓	✓	✓	✓	✓	✓	✓