Bring Your Own Data to X-PLAIN

Bring Your Own Data to x-Plain

Eliana Pastor
eliana.pastor@polito.it
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy

Elena Baralis
elena.baralis@polito.it
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy

ABSTRACT

Exploring and understanding the motivations behind black-box model predictions is becoming essential in many different applications. x-Plain is an interactive tool that allows human-in-the-loop inspection of the reasons behind model predictions. Its support for the local analysis of individual predictions enables users to inspect the local behavior of different classifiers and compare the knowledge different classifiers are exploiting for their prediction. The interactive exploration of prediction explanation provides actionable insights for both trusting and validating model predictions and, in case of unexpected behaviors, for debugging and improving the model itself.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Interpretability, Prediction Explanation, Local Rules

1 INTRODUCTION

Machine learning models are increasingly adopted to assist human experts in decision making. Especially in critical tasks, understanding the reasons behind model predictions is essential for trusting the model itself. Investigating model behavior can provide actionable insights. For example, experts can detect model wrong behaviors and actively work on model debugging and improvement. Unfortunately, most high performance models lack interpretability. To address this problem, we propose x-Plain, an interactive tool that allows human-in-the-loop inspection of classifier reasons behind predictions. The tool can help data scientists and domain experts to understand and interactively investigate individual decisions made by black box models. The demonstration will illustrate x-Plain key functionalities and scenarios. We will interactively engage the audience, inviting them to inspect and compare local behaviors of models trained on their own data sets.

The x-Plain interactive tool focuses on local interpretability for structured (i.e., tabular) data. Many techniques have been proposed for gaining local insights into black-box model behavior. A complete overview is presented in [2]. x-Plain leverages on LACE [5] as explanation method. It exploits local rules to provide qualitative local prediction interpretation, as performed by the explanation methods Anchor [6] and LORE [1], but, differently, it also quantitatively estimates the relevance of local rules and single attribute values in terms of prediction difference. The work [9] likewise analyzes the relevance of attribute values for a prediction by evaluating the prediction change if one or more attributes are jointly omitted. The information on attribute interaction is summarized in one global contribution for each attribute value. In the x-Plain system, the influence of each attribute value and significant sets of attribute values of a particular instance on the prediction of its class label is separately quantified.

Considering the relevance of human-in-the-loop inspection of model behavior, visual interfaces have been proposed. In [8], the authors propose a visual analytics interface that, differently from x-Plain, only works with binary classifiers and binary feature sets and enables interactive exploration of a set of instance-level explanations. Krause et al. propose the interactive visual analytic system Prospector [3]. It leverages on partial dependence plots as explanation method and it provides graphical representations of how features affect the predictions of a generic model. However, it relies on partial dependence for one attribute at a time. Hence, differently from x-Plain, it does not consider the influence that features jointly have.

2 X-PLAIN SYSTEM OVERVIEW

x-Plain [1] is an interactive tool that allows human-in-the-loop inspection of the decision-making process of machine learning models. x-Plain leverages on LACE [5], which is a model-agnostic explanation method to explain classifier predictions on single instances. LACE analyzes, by means of local rules, the relevance of each attribute value and significant sets of attribute values of a particular instance to provide the explanation of the prediction of its class label in terms of prediction difference [7].

Let f be an arbitrary trained classification model and x the instance whose prediction made by model f we want to

[1] Source code available at https://github.com/elianap/X-PLAIN-Demo, demo video at http://bit.ly/X-PLAIN-Demo-SIGMOD2020

explain. LACE [5] first step is the investigation of the locality of the particular prediction to be explained. The locality is captured by means of the K instances in the training set that are nearest to the instance x to be explained. Next, the K neighbors, labeled by the black box model itself, become training data for an interpretable associative classifier which extracts *local rules*. The local model is able to (i) provide a *qualitative* understanding by means of local rules and (ii) identify the (small) subset of attribute groups which are relevant for the prediction and are exploited for the *quantitative* evaluation of attribute importance. Finally, the prediction difference measures the prediction changes when one or more attribute values are omitted [5, 7]. Hence, it expresses a quantitative evaluation of the importance of (i) each attribute value, (ii) each relevant attribute subset derived by the local rules, and (iii) the union of all rule bodies.

The x-Plain tool has been developed in Python. The exploited explanation method works with discrete data [5]. Thus, continuous attributes are firstly discretized. Each data set is split into *training* and *explain* set. The training set is used to train classification models. Explanations are produced for instances in the explain set. x-Plain is model agnostic. Hence, it provides explanations and local inspection for individual predictions of any arbitrary classifier.

3 X-PLAIN DEMO SCENARIOS

The demonstration will show the effectiveness of x-Plain in providing insights on the multiple facets by which classification model behaviors may be analyzed. It will cover the following key functionalities of the x-Plain interactive tool.

Explanation of an instance prediction. x-Plain allows the evaluation of attribute value importance for the prediction of each class label, both for correct and mispredicted instances. This feature also enables the comparison of the local behavior for multiple target classes and classifiers.

Human-in-the-loop model analysis. Users may actively speculate and analyze their assumptions on the local model behavior based on their prior domain knowledge and perform what-if analysis by tweaking attribute values of single instances.

Explanation metadata analysis. The explanations provided by x-Plain provide actionable metadata that can be collectively exploited to characterize the global model behavior.

In the demonstration, artificial and real-world data sets from UCI repository [4] will be considered. The audience of the demo session will be invited to bring their own data and actively experiment with our tool. They will analyze and compare individual explanations and local behaviors of multiple classifiers, examining what models have learned locally for the newly proposed data set under analysis, in a "Bring Your Own Data" (BYOD) modality.

3.1 Explanation of an instance prediction

x-Plain generates explanations of instance x belonging to the *explain* dataset with respect to any arbitrary target class c. An explanation captures what model f has learned in the locality of x for class c in terms of local rules and prediction difference. A positive prediction difference indicates that the attribute value (or set of attribute values) has a positive influence on the target class label assignment. A negative one, instead, means that the attribute value(s) is against the assignment. In the following, a variety of explanation-based analyses is outlined.

Explanation of mispredicted instances. The x-Plain tool allows interactively inspecting the classifier behavior for misclassified instances. The explanation highlights the reasons why the classifier wrongly assigned the class label to a particular instance. The user can interactively select an incorrect prediction to inspect and target class c and the corresponding explanation is presented. Domain experts can inspect it and detect if the model has learned wrong associations. Hence, explanations (a) allow experts to comprehend why decisions are made, (b) enable model debugging and (c) foster model improvements in the case of model incorrect behaviors.

An example of misprediction inspection is presented in Figure 1a for the prediction of instance *y=platypus* of the *zoo* dataset made by a Naive Bayes (NB) classifier. The *zoo* data set belongs to the UCI repository [4]. The classification task is the identification of the biological class of animals, based upon its 16 variables. The NB classifier incorrectly assigns instance *y=platypus* to class *amphibian*. By exploiting x-Plain, users can inspect the reasons behind the wrong assignment. The extracted local rule is *{feathers=0, eggs=1, airborne=0, aquatic=1, predator=1, backbone=1, breathes=1, venomous=0, fins=0, legs=4, domestic=0}* $\longrightarrow$ *class='amphibian'*. The quantitative explanation, reported in Figure 1a, highlights that NB amphibian class assignment is driven by being an aquatic animal and laying eggs.

Explanation of correctly classified instances. The explanation of a correctly predicted instance occurs similarly to misclassifications. It highlights why the classifier has made that particular choice. Users can inspect the motivation behind the prediction. Hence, x-Plain supports GDPR compliant explanations by providing "meaningful information about the logic involved". Furthermore, the user can compare the explanation with her prior domain knowledge and determine if the model is "right for the right reasons".

Comparing the behavior for multiple target classes. Explanations of instance x prediction provided by the same model f for different target classes can be visually compared. Users can inspect and compare the subsets of attribute values that are critical and significant for each analyzed target class.

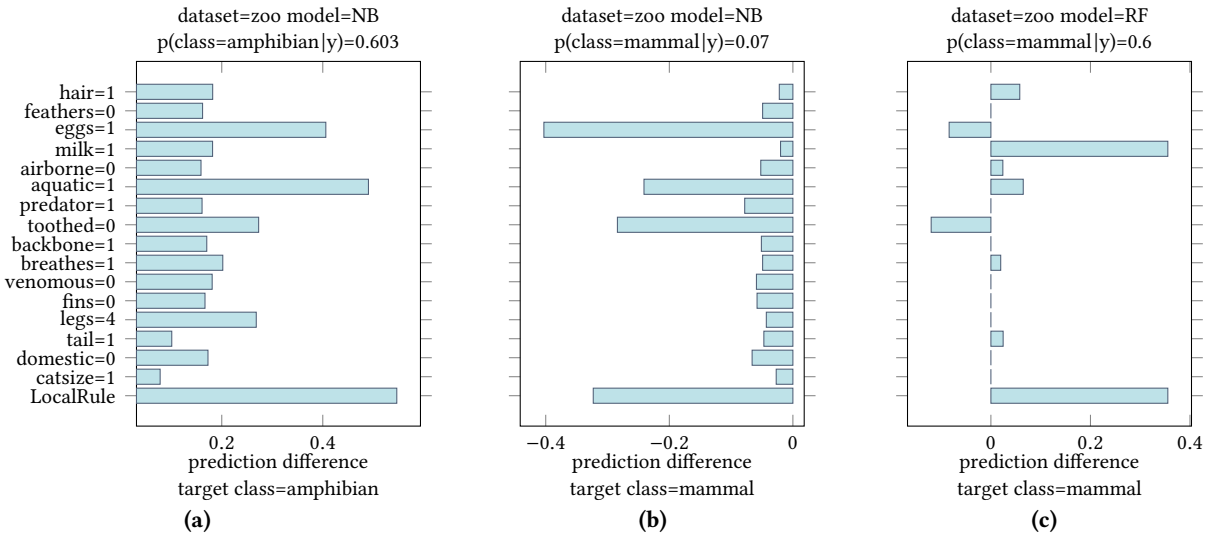

Figure 1: Explanations for instance *y=platypus* of the *zoo* data set for the NB prediction with respect to (a) *amphibian* and (b) *mammal* classes and (c) for the RF prediction with respect to *mammal* class.

The analysis is particularly interesting in case of misclassified instances. The explanation with respect to the true label highlights which attribute values have a negative influence on the true label assignment.

Consider again instance *y=platypus*. A user may be interested in investigating why the NB classifier does not assign instance y to the *mammal* class. The explanation, reported in Figure 1b, highlights as terms that have the most negative influence in the assignment to class *mammal* the characteristics of laying eggs and not being toothed. Hence, a user can carefully inspect the motivations for different classes and evaluate if the model under analysis indeed captures the distinguishing characteristics of the studied problem.

Comparing the behavior of multiple classifiers. Explanations of the same instance x made by different classifiers allow users to easily compare what the different models have learned. The comparison of the local behaviors may be exploited by experts to select the model that best fits a specific purpose. Users may also select which model prediction to trust based on their prior domain knowledge of the problem.

As example, we consider the prediction of instance y = *platypus* by a Random Forest model (RF). The RF model correctly identifies instance y as belonging to class *mammal*. Figure 1c shows the explanation of instance y for the predicted class. The local rule highlighted by x-Plain is *{feathers=0, milk=1, backbone=1, breathes=1, venomous=0}* $\rightarrow$ *class=mammal*. The term with the highest positive prediction difference is the animal characteristic of producing milk. On the other hand, being toothed and laying eggs have a negative influence for the mammal class assignment. Based on our knowledge of the biological class mammal we can say that RF has captured distinctive aspects of the class.

3.2 Human-in-the-loop model analysis

Human-in-the-loop inspections allow users to test their assumptions on the model internal behavior by actively modifying the classifier behavior as follows.

User rule definition. A user may interactively obtain the prediction relevance of additional, user-defined, rules. Based on prior domain knowledge, a user may expect a combination of attribute values to be important for the considered prediction. x-Plain directly estimates the prediction difference for the new user rule(s) and includes the new terms in the bar plot representation.

As an example, consider again the prediction of the NB model for instance *y=platypus*. We may be interested in investigating if NB, despite the wrong assignment, has learned some discriminant characteristics of the mammal class. Following the definition of the mammal biological class, we interactively define the new user rule *{milk=1, backbone=1, hair=1}*. x-Plain directly estimates the relevance of the subset, that is equal to 0.043. Hence, it shows that, in the NB model, this attribute value combination has a (small) positive influence on the mammal class assignment.

What-if analysis on attribute values. What-if analysis allows users to examine how and why the prediction of x could change if some of its attribute values were different. Users can interactively change the value of one or more attributes at a time. x-Plain directly provides the explanation of the prediction for the instance with the perturbed attribute values. Users can inspect the changes in (i) predicted class

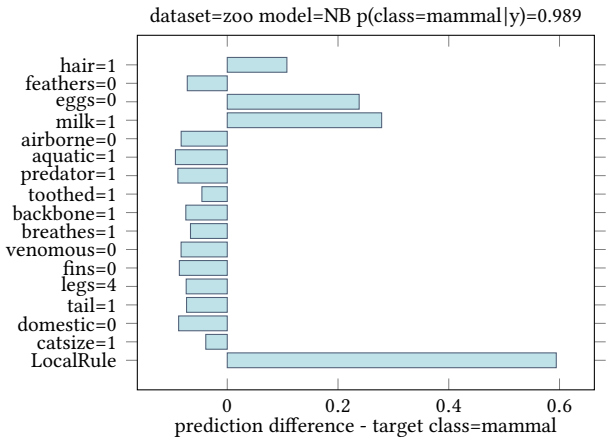

Figure 2: Explanation of tweaked instance *y* of the *zoo* data set for the NB prediction for mammal class.

label, (ii) local rules and (iii) prediction differences of the perturbed instance. Hence, they can explore model f labeling behavior when the attributes of interest are replaced with user-defined values.

Consider again instance *y=platypus* and the NB model. We analyze how the prediction and corresponding explanation would change if the two most discriminant negative terms highlighted by explanation in Figure 1b were different. We tweak the *eggs* and *toothed* attributes, setting them to 0 and 1 respectively. The resulting explanation computed for class *mammal* is reported in Figure 2 with local rule *{hair=1, feathers=0, eggs=0, milk=1, toothed=1, backbone=1, breathes=1, venomous=0}* $\rightarrow$ *'mammal'*. The perturbed instance is assigned to class mammal and the explanation shows that *milk=1* and *eggs=0* influence positively the prediction and the tweaked terms together interact for the class assignment.

3.3 Exploiting explanation metadata

Multiple local explanations generated by x-Plain may provide global insights on the model by highlighting which attributes and subsets of attribute values characterize the class assignment. Explanation metadata are generated by computing prediction explanations of model f for N instances of the *explain* dataset, considering as target class the predicted one, and stored in a knowledge base. Then, the average prediction difference is computed for each attribute value and subset of attribute values, separately for each target class. Finally, attribute value subsets are ranked based on average prediction difference. High-ranked combinations provide a description of the global model behavior for a given class.

As an example, consider the explanation metadata of the *zoo* dataset for a NB model and set N as the explain dataset cardinality. When selecting as target class the *mammal* one, x-Plain indicates that the most distinctive attribute is *milk*, attribute value *milk=1*, followed by the term *eggs=0* and subset of attribute values *{hair=1, feathers=0, eggs=0, milk=1, toothed=1, backbone=1, breathes=1, venomous = 0 }*. For the *bird* target class, the term *feather=1* is the most discriminant attribute value, followed by the term *legs=2* and *{ hair=0, feathers=1, eggs=1, milk=0, toothed=0, backbone=1, breathes=1, venomous=0, fins=0, legs=2, tail=1 }* is the most characterizing subset. Hence, by exploiting the metadata provided by a collection of prediction explanations, x-Plain may reveal which individual attribute values or subsets are overall most discriminating for each class.

ACKNOWLEDGMENTS

This work is partially funded by SmartData@PoliTO. We thank Andrea Cognolato for supporting the interface design.

REFERENCES

[1] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 93:1–93:42 pages.

[3] J. Krause, A. Perer, and K. Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *ACM CHI*.

[4] M. Lichman. 2013. UCI Machine Learning Repository.

[5] E. Pastor and E. Baralis. 2019. Explaining Black Box Models by Means of Local Rules *(SAC '19)*. 510–517.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*.

[7] M. Robnik-Šikonja and I. Kononenko. 2008. Explaining Classifications For Individual Instances. *IEEE TKDE* 20, 5 (2008), 589–600.

[8] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. In *ACM HILDA*.

[9] E. Štrumbelj and I. Kononenko. 2010. An Efficient Explanation of Individual Classifications Using Game Theory. *JMLR* 11 (2010), 1–18.