Received 26 April 2016; | Revised 6 June 2016; | Accepted 6 June 2016

DOI: xxx/xxxx

ARTICLE TYPE

Visual Diagnostics of a Model Explainer – Tools for the Assessment of LIME Explanations

Katherine Goode*[1] | Heike Hofmann[1,2]

[1]Department of Statistics, Iowa State University, Iowa, United States
[2]Center for Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, Iowa, United States

Correspondence
*Corresponding author. Email: kgoode@iastate.edu

Present Address
This is sample for present address text this is sample for present address text

Summary

This is sample abstract text.

KEYWORDS:
LIME, black-box models, interpretability, diagnostics

To check on when editing writing:

- Generally: avoid 'can be'. Replace by 'is'.
- References like 'they', 'them' ... replace them by repeating the noun you are referring to to avoid any kind of ambiguity
- words to go back and make sure I am not using too often: ability, produce, understand

1 | INTRODUCTION

*** I added some subsections to the introduction, because it was getting so long. What do you think? Would it be better to make the introduction shorter and move the details to a separate section?

In the field of statistics, there are two main uses for models: inference and prediction. Machine learning models are often used for the latter purpose. While these models have been proven to perform well in a wide range of prediction problems, their accuracy comes at the cost of interpretability. Machine learning models are generally complex algorithms that are good at identifying patterns in the data but often lack a functional form, which makes it impossible to directly interpret them ***Should I change this them?. As a result, machine learning learning models produce predictions that lack an explanation, which has earned them ***What about this one? the reputation of being "black-box models". HH: I would make the claim that the need for explanations goes even deeper. Even if a functional form exists, such as in a linear model, there are situations where the model gets complicated enough to be not or not easily interpretable. Only very good natured three-way interactions in a logistic regression are interpretable. For the other ones more work is needed than most audiences would be capable of. Explainer models can be used in those situations as well.

1.1 | The Importance of Explanability

The ability to interpret a model serves multiple purposes. When a model is first fit, interpretations help to diagnose the model. Knowing which variables influence a prediction makes it possible to determine if the predictions are based on reasonable variables. If not, appropriate adjustments to the model can be made. After a model has been diagnosed, the explanations for the predictions are used to understand the underlying mechanism that produced the data and provide an argument for why the predictions should be trusted. In some areas of application, this ability to be able to explain the results from a model is critical.

The forensics sciences and the heath care industry are two areas that have benefited from machine learning but have an obvious need for explainable predictions. For example, Hare et al. [2] discuss the use of a random forest model in the forensics sciences to determine whether two bullets were fired from the same gun. Yu et al. [5] describe how healthcare has advanced through the use of machine learning in ways such as using neural networks to automate medical image diagnoses and implementing Bayesian networks to predict clinical outcomes. In both of these fields, the decisions made based on the results of the machine learning models can greatly impact people's lives. If it is not possible to explain the output from a model, it becomes questionable whether to rely on the predictions to make important decisions.

As an example, suppose that forensics examiners are trying to determine if a bullet used in a crime was fired from the gun that belongs to a suspect in the investigation. A random forest model could be used to produce a probability that the bullet was fired from the gun under question. Even if the model returns a high probability of a match, it will be difficult for a jury to trust the results from the model when deciding whether to convict the suspect without the forensics examiners having the ability to explain which factors in the model led to the high probability.

The European Union has taken the desire to provide explanations for machine learning model predictions a step further. In May 2018, the General Data Protection Regulation (GDPR) went into effect [1]. The regulation includes a policy that gives the right for individuals affected by decisions made via automated algorithms to understand the logic behind the decisions being made. As Goodman and Flaxman [1] point out, this regulation has a great effect on the way that machine learning algorithms are used to make decisions. The wording of the policy leaves room for interpretation, but the authors believe that, at a minimum, the regulations will require that an explanation of how the input variables relate to the predictions be provided.

1.2 | Explainer Models

As a result of the movement to produce explainable predictions, an area of research has emerged focusing on developing ways to explain output from machine learning algorithms. One approach that is being taken is to develop model explainers that provide insight into the performance of the complex model (reference). A model explainer is a method that is separate from the model but uses the model and output from the model to shed light on the process that the model goes through to produce predictions. LIME (local interpretable model-agnostic explanations) is one such model explainer [4].

While some model explainers are focused on understanding a model at the global level, LIME is designed provide a local explanations by focusing on a single prediction of interest. Additionally, LIME was designed to work with any model and to produce easily understandable explanations. [4] XXX a citation might be good here for local and agnostic Is it okay to include the same reference to the original LIME paper again? sure Conceptually, LIME fits a simple interpretable model, referred to as the explainer, that is meant to capture the behavior of the complex "black-box model" in a local region around a prediction of interest. The interpretation of the simple model is then used to explain the variables that most influenced the prediction made by the complex model.

I think it would make sense to combine figures 1 and 2 into one big, two-column spread figure with the two explainer models shown side by side. Include a legend in the plot for the size of points. It would also be nice to change the points from black to something a bit lighter - maybe grey30? - and add a bit of alpha blending to them. That will allow you to still see the overall structure, but show the underlying 2d density somewhat. It also increases the contrast to the black lines of the explainer model. Figure 1 provides a visualization of this conceptual understanding of LIME. These figures show the predictions from a hypothetical black-box model plotted against the two features used to fit the black-box model. The orange point It's better not to refer to colors in the text. Why don't you double encode with shape and refer to the shape of that special point. We do not know how color is going to be publicized and it's painful to find all the textual references. represent one prediction that is of interest to explain. The size of the points represent the proximity to the prediction of interest measured using the Gower distance metric (reference and add exponent value?) yes. The black lines represent the explainer model which in this case is a weighted linear regression model fit with the black-box predictions as the response variable and Features 1 and 2 as the covariates in the model with proximity values used as weights in the model. The top image shows that there is a complex relationship between black-box predictions and Feature 1. Here the explainer model is plotted with Feature 2 set to be the observed value of Feature 2 for the prediction of interest. The explainer model captures the relationship in the local region around the prediction of interest well. The bottom

[1] https://eugdpr.org/the-regulation/gdpr-faqs/

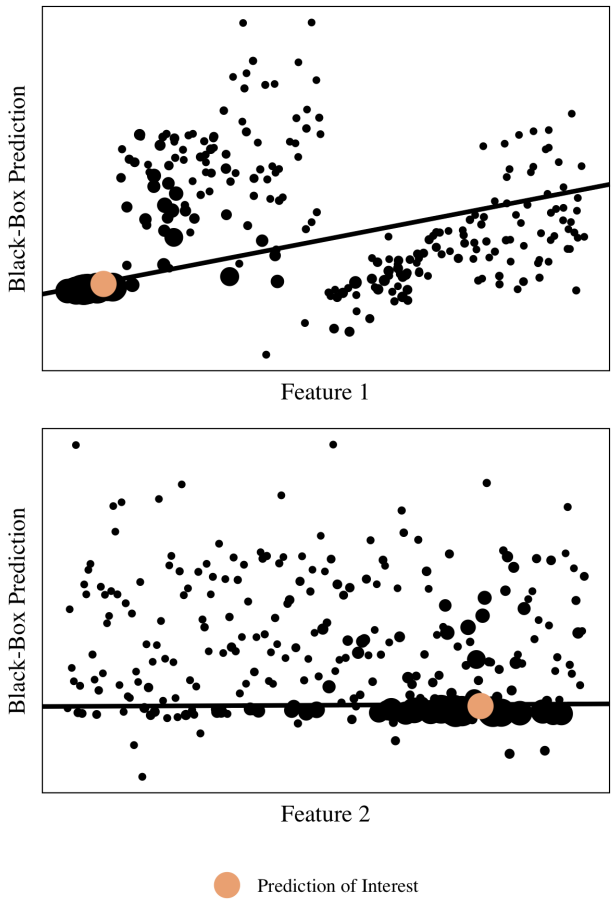


FIGURE 1 Conceptual depiction of LIME. The black lines represent a weighted linear regression model being used as the explainer model for a hypothetical black-box model fit with two features. The explainer model suggests that Feature 1 is driving the prediction made by the black-box model for the case of interest since it has captured the slope in the local region around the prediction of interest.

figure shows that there is no relationship between the black-box predictions and Feature 2 in either the global sense or in a local region around the prediction of interest. Here, the explainer model is plotted with Feature 1 set to be the observed value of Feature 1 for the prediction of interest, and it has a slope of approximately 0. Taken together, this explainer model would indicate that Feature 1 is driving the prediction made by the black-box model for the prediction of interest.

1.3 | Motivation for Diagnosing LIME

While reading through the original paper again, I rediscovered this quote. This may be something to reference or quote in this section: "...our choice of G (sparse linear models) means that if the underlying model is highly non-linear even in the locality of the prediction, there may not be a faithful explanation. However, we can estimate the faithfulness of the explanation on $\mathcal{Z}$, and present this information to the user. This estimate of faithfulness can also be used for selecting an appropriate family of explanations from a set of multiple interpretable model classes, thus adapting to the given dataset and the classifier. We leave such exploration for future work, as linear explanations work quite well for multiple black-box models in our experiments."

At a conceptual level, explainer models add another level of complexity to predictive models: in trying to explain the black-box model, a simpler explainer model is added. To be able to trust in the explanation it is imperative to check that the explainer model is reliable and does not over-simplify the black-box model.

The current implementations of LIME use a linear regression model as the explainer model [3] (references to R and Python packages) for which models? LIME uses neural networks for image analyses I think that it is actualy using ridge regression for all cases - even images. These implementations are relying on the assumption that the relationship between the complex model predictions and the features is linear in a local region. HH: be a bit careful in the phrasing here - we can always approximate any function using a linear form (think Taylor expansion). The question is how big the error is. It is important to assess this assumption in the local region of interest in order to trust the explanations produced by LIME. The next sentence is muddying the waters - by being able to doagnose explainer models we will be able to assess the different settings offered in the implementations. Let's bring those up later. Additionally, the implementations offer various input settings, but little work has been done to provide advice on how to specify the settings in practice (look more into this). An assessment of the explanations could also help to determine which settings produce the explainer model with the best approximation of the complex model.

Figure 2 provides an example where the explainer model is doing a poor job of approximating the complex model. The plots were created using the same data as Figure 1 , but the Gower distance metric was adjusted (add exponent value), so that observations further away from the prediction of interest are given larger weights. In this case, the explainer model is doing a poor job of

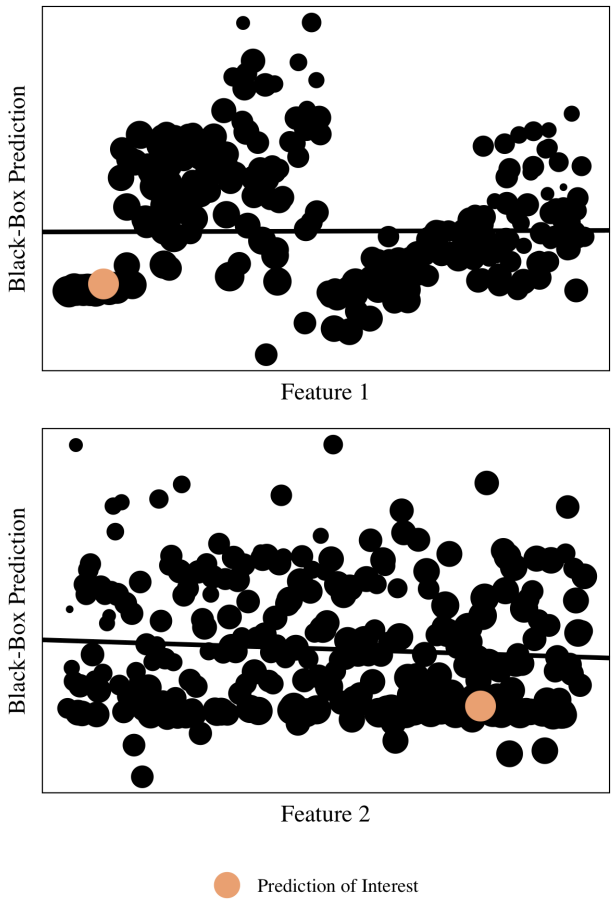


FIGURE 2 A second version of explainer model applied to the data from Figure 1 . The model explainer is refit using an adjusted distance measure. It leads to an explainer model that is doing a poor job of approximating the complex model in the local region around the prediction of interest for both features.

capturing the relationship between the black-box predictions and the feature for both Feature 1 and Feature 2. However, the magnitude of the slope of the line in the bottom figure appears to be larger than . As a result, this explainer model would return an explanation that Feature 2 is driving the prediction made by the black-box model for the case of interest, which is not an accurate explanation.

In the scenario depicted in Figure 1 , the distance metric was adjusted until the explainer models produced a good local approximation of the complex model. HH: excellent! What are the respective R squares? For Figure 2 , the default setting for distance measures was used by the LIME R package. Without these images, it would not be obvious that the explainer model used in Figure 2 is a poor local approximation.

1.4 | Overview of Paper

In this paper, we will present some visualizations tools to assess the explanations from LIME. While predictive models are used in both regression and classification settings, we will focus on the classification setting for this paper. For additional simplicity, we will only discuss the case with a dichotomous response variable, but we believe that the work is adaptable for other situations.

Section 2 provides some background on the LIME algorithm and the current implementations of LIME. In Section 3, we discuss ways to assess the LIME explainer model and introduce our visualization tools. Section 4 demonstrates the use of our diagnostic tools with a random forest model fit to a forensics bullet matching dataset (***I went ahead and excluded the idea of including the iris data). To conclude, Section 5 reviews the importance of assessing the LIME explainer model, discusses how the results from the example provide possible insights into the workings of LIME, and suggests future research directions.

2 | BACKGROUND ON LIME

LIME was developed in 2016 by Ribeiro et al. [4]. The authors were interested in developing a model explainer method that produced easily understandable explanations for individual predictions made by any predictive model [4]. It was initially implemented in a Python package[2] by the original authors and was later adapted to an R package by Thomas Lin-Pedersen[3].

The LIME algorithm is presented in the original paper in a general context that accounts for cases such as text classification and feature recognition. For this paper, we are only considering the tabular data situation with a binary categorical response variable and continuous feature variables. As a result, the following subsections are written in terms of this context. Furthermore, in this paper we are relying on the LIME implementation in R [3]. This implementation deviates at times from the original implementation by Ribeiro et al. [4]. Whenever the implementations differ, we will highlight these deviations.

[2]https://github.com/marcotcr/lime
[3]https://github.com/thomasp85/lime

2.1 | LIME Procedure in the Context of Tabular Data with a Dichotomous Response Variable

***Need to clean up the notation. Should I use the same notation as the original paper? Let $\mathbf{X}$ be an n by p data matrix with p features and n observations, and let $x = \left(x_1 \; x_2 \; \dots \; x_p \right) \in \mathbb{R}^p$ be the observation of interest. Furthermore, let $\mathbf{y}$ be a vector of length n of response values and $y \in \{0, 1\}$ be the observed response value associated with x. Suppose that f is a classification model where $f : \mathbb{R}^p \to [0, 1]$ that is applied to $\mathbf{X}$ and $\mathbf{y}$. Let the predictions made by f applied to $\mathbf{X}$ be denoted as $\hat{\mathbf{y}}$. It is of interest to explain the prediction made by f when f is applied to x. Note that $f(x)$ is equal to the probability that $y = 1$. Given this setup, the LIME procedure is as follows.

1. Generate a new dataset $\mathbf{X}^*$ of size m by p using the observed values in $\mathbf{X}$. There are various ways to simulate the new dataset, and the methods currently used by the LIME R package will be described in more detail in Section 2.2.

2. Apply f to $\mathbf{X}^*$ to obtain a vector of predictions $\hat{\mathbf{y}}'$ of length m. Let the prediction made by f when applied to the transformed case of interested be denoted as $f(x') = \hat{y}'$.

3. Apply a transformation T to $\mathbf{X}^*$ that results in an interpretable representation. The transformation that is applied will depend on the simulation method used. Section 2.2 will discuss the transformations used by the LIME R package. Let $T(\mathbf{X}^*) = \mathbf{X}'$. Furthermore, apply T to x to obtain $T(x) = x'$.

4. Compute a proximity measure between x and $\mathbf{x}'_i$ for each $i \in 1, ..., m$ denoted by $\pi_x(\mathbf{x}'_i)$.

5. Identify a class of potentially interpretable models such as linear models or decision trees. Denote this class of models by $G : \mathbb{R}^p \to [0, 1]$. Section 2.2 describes the class of interpretable models used by the R package.

6. Perform feature selection...

7. Fit explainer model

8. Interpret

Somehow I need to mesh the procedure peformed by the R package and the description of LIME described in the original paper. The place where I left off in the procedure above is where the two start to disagree. Here are my old "notes" on the LIME algorithm based on the paper:

- G: class of potentially interpretable models (e.g. linear models, decision trees, rule lists) in our case, we will use ridge regression
- g: explanation model where $g : \{0, 1\}^{p'} \to \mathbb{R}$ and $g \in G$ do we need $\{0, 1\}^{p'}$?
- $\Omega(g)$: measure of complexity of g (e.g. depth of a tree, number of non-zero coefficients in a linear model fit using LASSO)
- $\mathcal{L}(f, g, \Pi_x)$: the fidelity functions which is a measure of how unfaithful g is in approximating f in the locality defined by π_x
- $\xi(x)$: explanation produced by LIME where

$$\xi(x) = \underset{g \in G}{\arg\min} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

(i.e. want to minimize $\mathcal{L}(f, g, \Pi_x)$ and keep $\Omega(g)$ low enough to be interpretable by humans)

XX references :)

2.2 | Implementation of LIME in R

make sure to include an output figure of a LIME explanation here

3 | METHODS

3.1 | Diagnostic Tool 1

3.2 | Diagnostic Tool 2

3.3 | etc...

4 | APPLICATION

4.1 | Bullet Matching Data

5 | DISCUSSION

ACKNOWLEDGMENTS

This is acknowledgment text. Provide text here.

Author contributions

This is an author contribution text. This is an author contribution text. This is an author contribution text. This is an author contribution text. This is an author contribution text.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

How to cite this article: Goode K., H. Hofmann, 2019, Visual Diagnostics of a Model Explainer – Tools for the Assessment of LIME Explanations, *Stat Anal Data Min: The ASA Data Sci Journal*, *volume, number and page.*

APPENDIX

A SECTION TITLE OF FIRST APPENDIX

References

[1] Goodman, B. and S. Flaxman, 2016: European Union regulations on algorithmic decision-making and a "right to explanation". doi:10.1609/aimag.v38i3.2741.

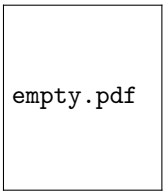
[2] Hare, E., H. Hofmann, and A. Carriquiry, 2016: Automatic Matching of Bullet Lands. *Annals of Applied Statistics*, doi:http://adsabs.harvard.edu/abs/2016arXiv160105788H.

[3] Pedersen, T. L. and M. Benesty, 2018: *lime: Local Interpretable Model-Agnostic Explanations*. R package version 0.4.0.
URL https://CRAN.R-project.org/package=lime

[4] Ribeiro, M. T., S. Singh, and C. Guestrin, 2016: "why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144.

[5] Yu, K.-H., A. L. Beam, and I. S. Kohane, 2018: Artificial intelligence in healthcare. *Nature Biomedical Engineering*, **2**, 719, doi:10.1038/s41551-018-0305-z.

AUTHOR BIOGRAPHY



Author Name. This is sample author biography text this is sample author biography text