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. They are often complex algorithms that are good at identifying patterns in the data but lack a functional form *** Is it okay for me to make this statement about the functional form?, so it not possible to directly interpret them. As a result, they produce predictions that lack an explanation, which has earned them the reputation of being "black-box models".

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? 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.

Figure 1 provides a visualization of this conceptual understanding of LIME. These figures show the predictions from a hypothetical black-box model plotted against

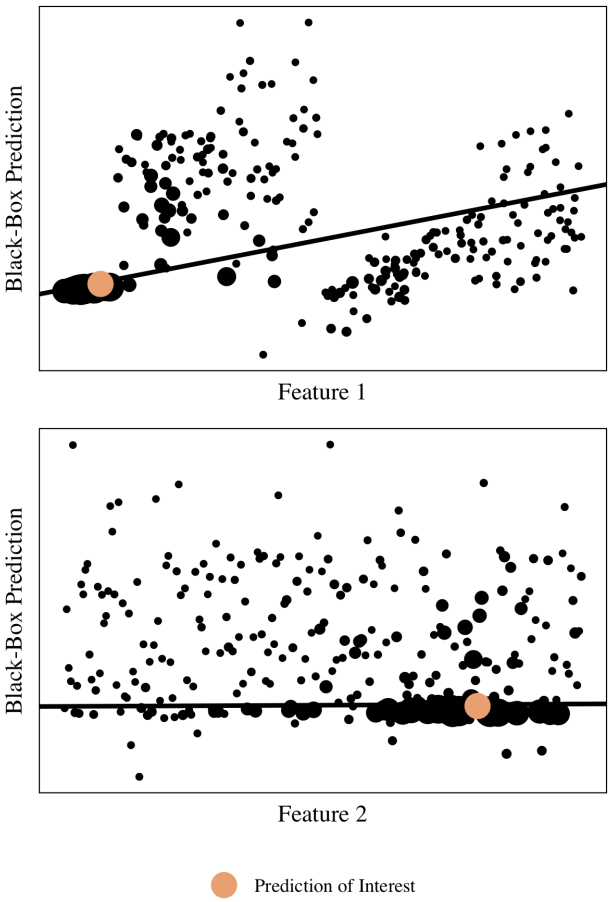

FIGURE 1 Conceptual depiction of LIME. The black lines represent a weighted linear regression model being used as the explainer model for a hypothetical black-box model fit with two features. The explainer model suggests that Feature 1 is driving the prediction made by the black-box model for the case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

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

the two features used to fit the black-box model. The orange point 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?). The black lines represent the explainer model which is a weighted linear regression model in this case fit with the black-box predictions as the response variable and Feature 1 and Feature 2 as the covariates in the model where the proximity values used as the 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 does a good job of capturing the relationship in the local region around the prediction of interest. The bottom shows that there is no relationship between the black-box predictions and Feature 2 in either the global sense or 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. This explainer model would indicate that Feature 1 is driving the prediction made by the black-box model for the prediction of interest.

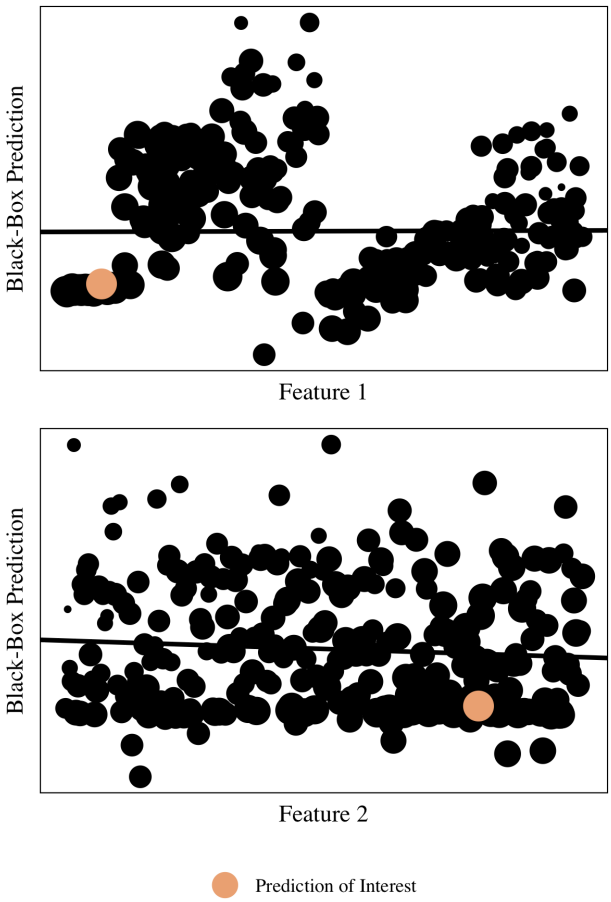

FIGURE 2 A second version of explainer model applied to the data from Figure 1 . The model explainer is refit using an adjusted distance measure. It leads to an explainer model that is doing a poor job of approximating the complex model in the local region around the prediction of interest for both features.

1.3 | Motivation for Diagnosing LIME

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 (references to R and Python packages). These implementations are relying on the assumption that the relationship between the complex model predictions and the features is linear in a local region. It is important to assess this assumption in the local region of interest in order to trust the explanations produced by LIME. 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 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. 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. 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 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 $T(x) = x'$. ***I need to check if the predictions are made before or after the transformation.

3. 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 $f(x') = y'$.

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. Apply the class of models G to $\mathbf{X}$...

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 :) Features from the training data are used to simulate a new dataset on which the simple model is fit. more on how the data is simulated - that's an important step. The complex model is then applied to the simulated dataset to obtain predictions. The observations associated with predictions are used as the response variable in a ridge regression model with the simulated features as the predictor variable with the highest weight given to observations closest to the prediction of interest. more on the weights Feature selection is performed to identify the most important variables in the local region. A final ridge regression model is fit with the selected features, and the coefficients of the model are used to interpret the behavior of the complex model.

2.2 | Implementation of LIME in R

The LIME R package allows for the following four methods to sample the perturbations based on the distributions of the features from the training data.

- Equally Spaced Bins
- Quantile Bins
- Normal Approximation
- Kernel Density Approximation

The methods of equally spaced bins and quantile bins also allow the user to specify the number of bins. As of now, there are no recommendations or procedures provided for how to determine which method to use. By default, LIME uses four quantile bins. It was of interest to see how the explanations from LIME varied across the four sampling methods when applied to the bullet matching data. The LIME algorithm was applied to each prediction from the test data obtained from the 'rtrees' random forest model for each of the four sampling methods. Within the bin based sampling methods, the algorithm was applied for 2 to 6 bins. It was decided to only go up to 6 bins since the more bins used the more complex the explanation becomes.

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

3 | METHODS

Ways to understand if the LIME explainer is doing a good job:

- diagnose the explainer to make sure that it is fitting the data well
- compare the prediction made by the explainer to the complex model prediction
- visualize all explanations to understand if the explanations are local or global
- compare results from different input options

3.1 | Diagnostic Tool 1

3.2 | Diagnostic Tool 2

3.3 | etc...

4 | APPLICATION

4.1 | Bullet Matching Data

Of particular interest during this assessment are the cases when the model is wrong.

1. explainer model generally has very low R^2 (probably due to binning) 2. "local" explanations are not local but are driven by the ("global") marginal distributions of covariates

In order to assess the LIME explanations created using different sampling methods, it was of interest to compare the top three features chosen as the important predictors by lime within a case from the test data across the different sampling methods. Figure ... is a heat map showing the top feature chosen by lime for each of the cases in the test data and different bin based sampling methods. The rows represent the cases in the test data, and the columns

represent the sampling methods. There are twenty methods included in the plot. These include the equally spaced bins and quantile bins from the lime package and the random forest score tree based bins and the same source tree based bins proposed in this paper. The rows are faceted by the test set and whether or not the observation is a match or not. The columns are faceted by these methods, and the columns within a facet represent the number of bins. Each method has 2 to 6 bins. The colors represent the top feature chosen by lime.

The variables of ccf and cms immediately show up as common variables chosen across all of the sampling methods. However, the patterns across the number of bins withing the sampling methods are different. When equally spaced bins are used, the top feature chosen is consistent across all cases within a number of bins category. For example, ccf is almost always chosen (change to actual number) with 2 equally spaced bins, matches is almost always chosen with 3 equally spaced bins, and non_cms is always chosen for the nonmatches with 5 and 6 equally spaced bins. This shows that with the bullet matching data, the top feature chosen with equally spaced bins is an artifact of the number of bins used. With equally spaced bins, this figure suggest that LIME is providing global explanations as opposed to local explanations. It would be preferable that the top feature chosen was more consistent across the number of bins and more variable across the cases. This would suggest that the top feature chosen is dependent on the feature values associated with a particular case and not just on which feature is the best explainer when b number of bins are used.

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

empty.pdf

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