# Visual Diagnostics of a Model Explainer – Tools for the Assessment of LIME Explanations from a Random Forest Model

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

[1]Department of Statistics, Iowa State University, Iowa, United States
[2]Center for (CSAFE, Iowa State University, Iowa, United States

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

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

**Summary**

This is sample abstract text.

**KEYWORDS:**
LIME, random forest, interpretability, diagnostics

## 1 | INTRODUCTION

Machine learning algorithms have been proven to provide accurate predictions. However, this accuracy comes at a cost. With complicated algorithms, the ability to understand the mechanism in the data and the functional form of the model are lost. This has led for a desire to be able to interpret machine learning algorithms. One approach to remedy this problem would be to develop a model explainer. The objective of such an explainer is to understand and explain model performance.

LIME (locally interpretable model explanations) is one such model explainer that was developed for this purpose (reference). Conceptually, LIME

- LIME does... - conceptually: models at two levels 1. Simple = explainer model 2. Complex = original "black box model" - usually: model predictions, maybe with ground truth - type I error - type II error - model is wrong in both cases - explanation is also a prediction - how reliable is that explanation? - we will present some tools that could help answer that question - we will show an example using this tools (bullet data) - we will mention some possible insights about LIME that these tools have shown us 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 - Describe LIME including details on binning and linear regression in binned features and motivation for the binning

## 2 | DATA

## 3 | METHODS

### 3.1 | LIME Algorithm

**Notation**

- $x \in \mathbb{R}^d$: original representation of an instance being explained

- e.g. feature vector containing word embeddings

- $x' \in \mathbb{R}^{d'}$: vector for the interpretable representation of the instance being explained

    - e.g. bag of words

- $G$: class of potentially interpretable models

    - e.g. linear models, decision trees, rule lists

- $g$: explanation model where $g : \mathbb{R}^{d'} \to \mathbb{R}$ and $g \in G$

- $\Omega(g)$: measure of complexity of $g$

    - e.g. depth of a tree, number of non-zeros in a linear model

- $f$: model that is being explained where $f : \mathbb{R}^{d} \to \mathbb{R}$

    - note: in classification $f(x)$ is the probability that $x$ belongs to a certain class

- $\Pi_x(z)$: proximity measure between an instance $z$ to $x$ which defines a locality around $x$

- $\mathcal{L}(f, g, \Pi_x)$: a measure of how unfaithful $g$ is in approximating $f$ in the locality defined by $\Pi_x$

**Step by Step Procedure**
Step 1:

## 3.2 | Application of LIME to Bullet Matching Data

### 3.2.1 | lime R Package Perturbation Creation Methods

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.

### 3.2.2 | Proposed Bin Creation Methods
### 4 | RESULTS

## 4.1 | LIME Package Explanations Dependent on Sampling Method

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 diferent sampling methods. Figure ... is a heatamp 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 facetted 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 sugggest 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.

## ACKNOWLEDGMENTS

## 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:
**Figure S1.** 500 hPa geopotential anomalies for GC2C calculated against the ERA Interim reanalysis. The period is 1989–2008.
**Figure S2.** The SST anomalies for GC2C calculated against the observations (OIsst).

**How to cite this article:** Goode K., H. Hofmann, (2016), A regime analysis of Atlantic winter jet variability applied to evaluate HadGEM3-GC2, *Q.J.R. Meteorol. Soc., 2017;00:1–6.*

# APPENDIX

## A SECTION TITLE OF FIRST APPENDIX

### References

[1] Elbaum, S., A. G. Malishevsky, and G. Rothermel, February 2002: Test case prioritization: a family of empirical studies. *IEEE Transactions on Software Engineering*, **28**, no. 2, 159–182, doi:12345.2345.

## AUTHOR BIOGRAPHY

empty.pdf

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