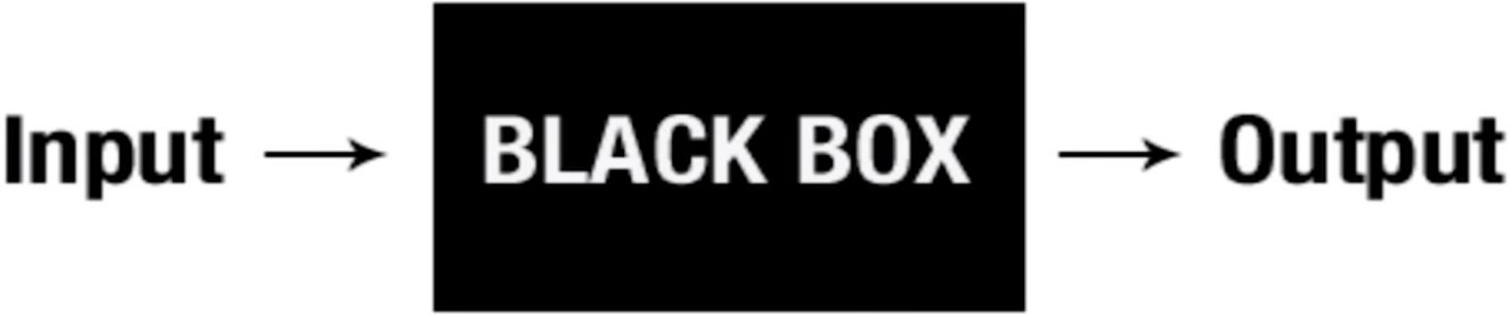
Do you trust your model?
An introduction to LIME

Fatima Loumaini

linkedin.com/in/fatima-loumaïni85

Why should I trust you?



What is LIME?

In 2016, Marco Tulio Riberio, Sameer Singh, and Carlos Guestrin from UW released a paper, “Why Should I Trust You?: Explaining the Predictions of Any Classifier” that was discussed at KDD 2016 as well as within a blog post. Within the paper, Riberio, Singh, and Guestrin propose LIME as a means of “providing explanations for individual predictions as a solution to the ‘trust the prediction problem’, and selecting multiple such predictions (and explanations) as a solution to ‘trusting the model’ problem.” Riberio, Singh, and Guestrin also define LIME as “an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model”.

What is LIME?

Local: refers to how we get to these explanations. It approximates the black box locally and the neighborhood of the explanations being explained.

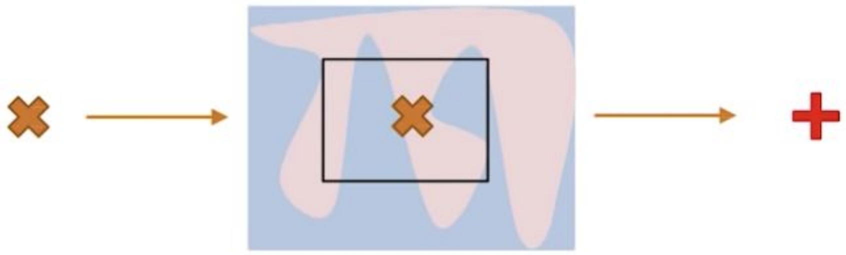
Interpretable: Explanations produced by Lime can be interpreted and understood by a human.

Model-Agnostic: Treats the model as a black box and it works for any model.

Explanations: Understand what the model does and which features it picks from to develop the predictions.

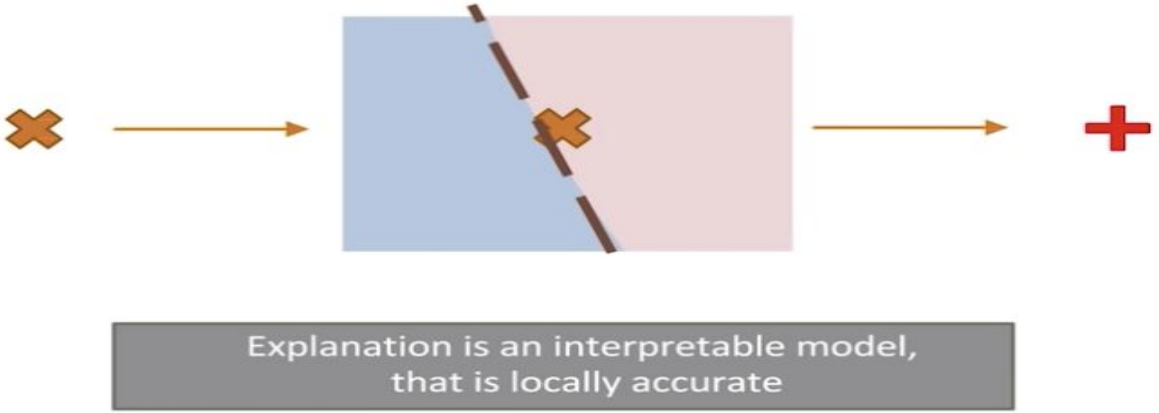
Being Local and Model Agnostic

Let’s pretend this is the landscape we have between predictors and outcomes variables and we don’t get much by having a global view.



Being Local and Model Agnostic..

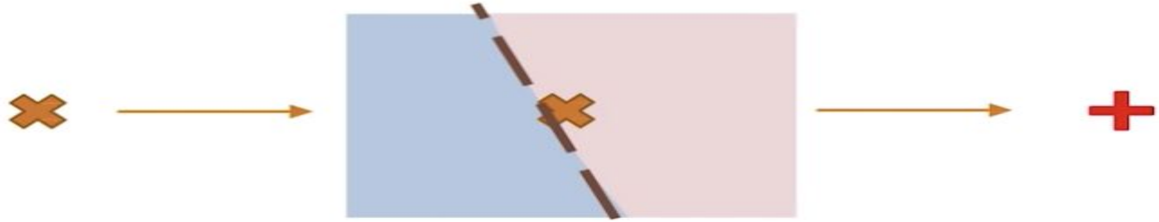
What LIME does, it ignores that and it goes very locally to a class of observation or an observation and gets to the point that it can use a linear regression to make a prediction and it will be very accurate at that level.



Explanation is an interpretable model, that is locally accurate

How Lime works

It makes use of the fact that linear models are easy to explain because they are based on linear relationships between features and class labels: The complex model function is approximated by locally fitting linear models to permutations of the original set.

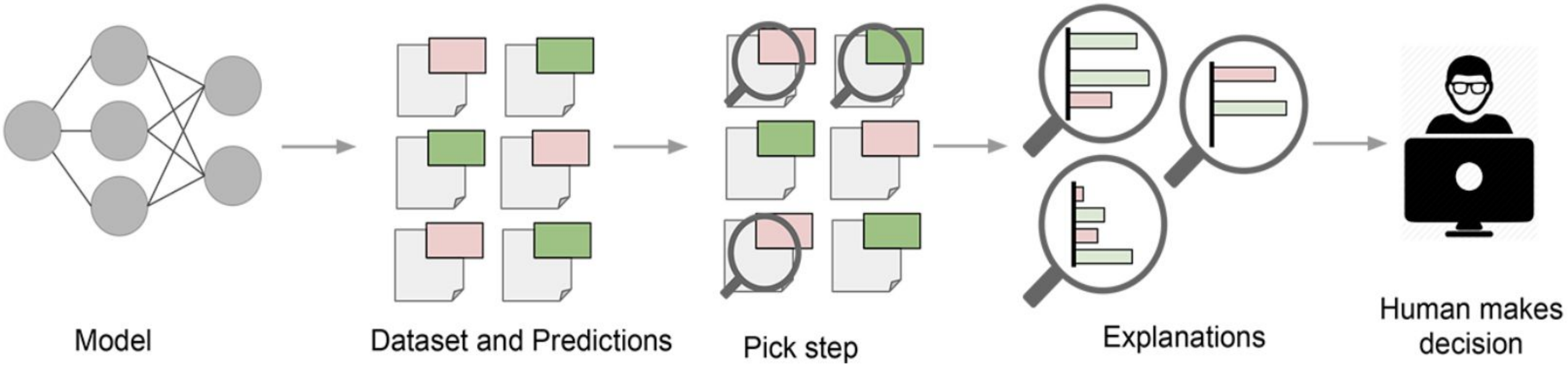


Explanation is an interpretable model, that is locally accurate

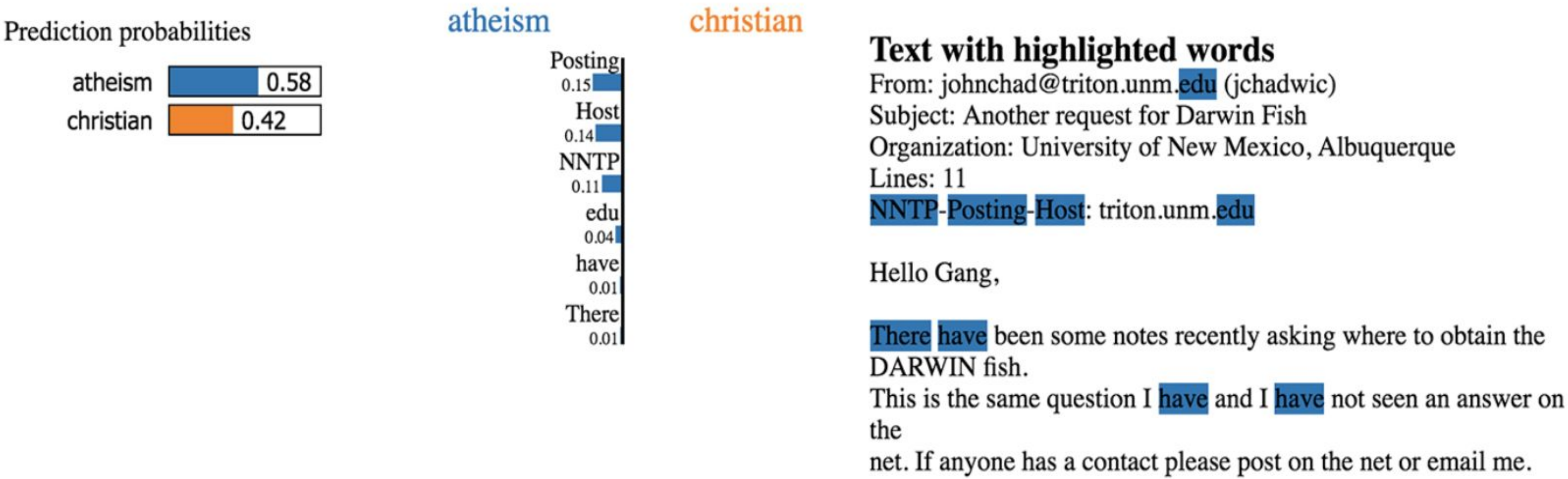
How Lime works

On each permutation, a linear model is being fit and weights are given so that incorrect classification of instances that are more similar to the original data are penalized (positive weights support a decision, negative weights contradict them). This will give an approximation of how much (and in which way) each feature contributed to a decision made by the model.

LIME in practice



LIME Examples



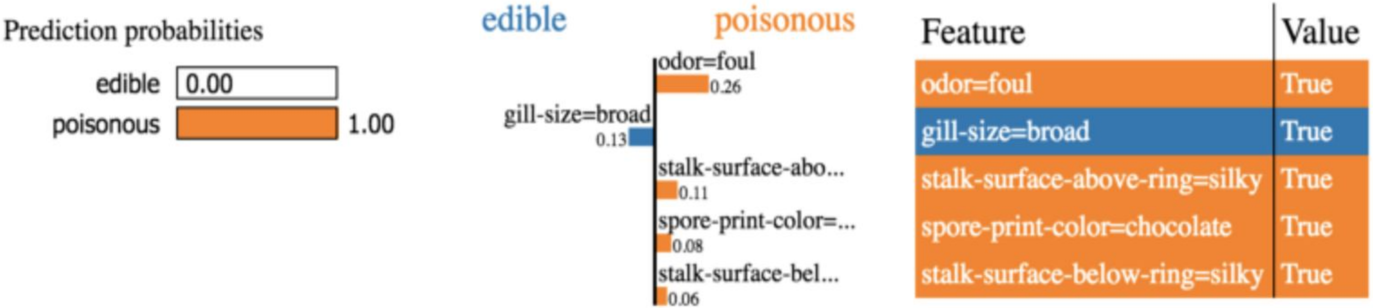
Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

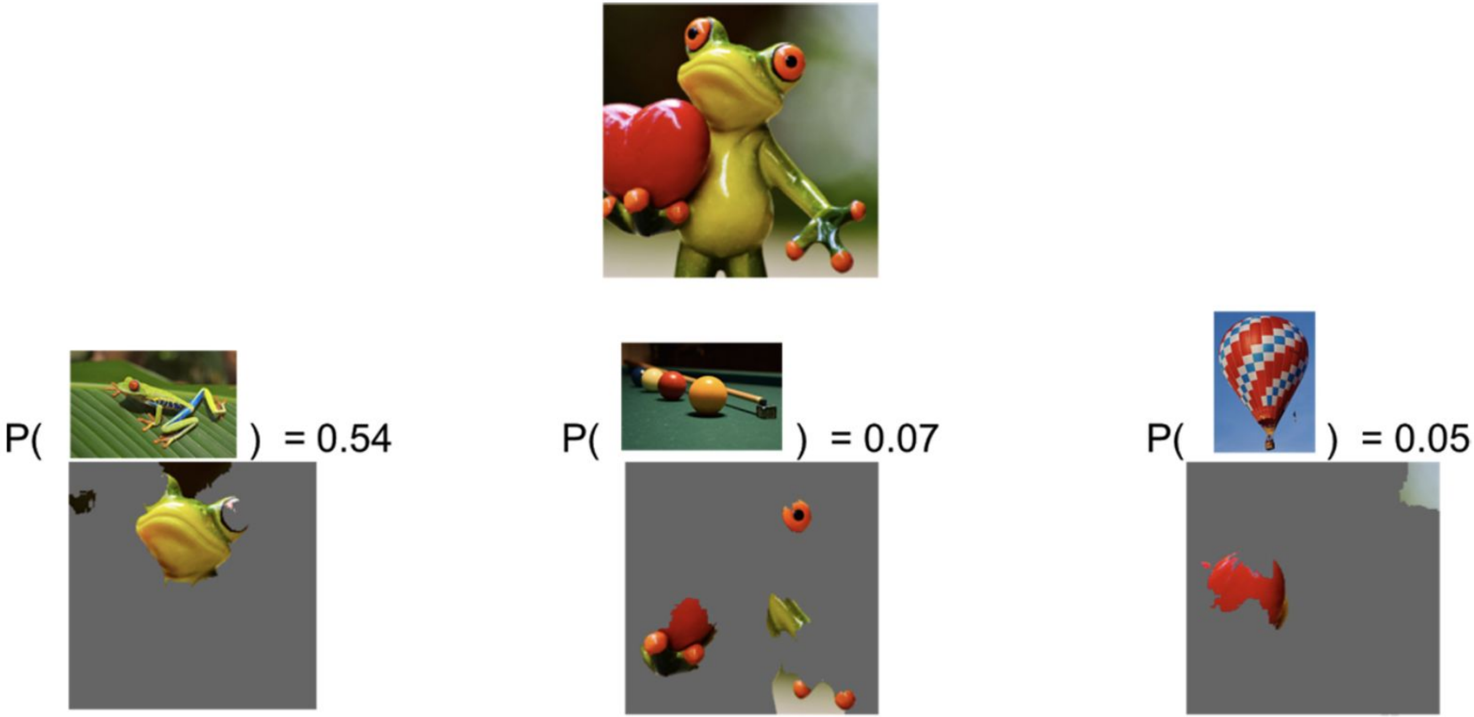
There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

LIME Examples



Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

LIME Examples



Why we should use LIME?

Trust is crucial for effective human interaction with machine learning systems, and explaining individual predictions is an effective way of assessing trust. LIME is an efficient tool to facilitate such trust for machine learning practitioners and a good choice to add to their tool belts (available in Python and R).

On a legal side, in Europe for example companies that incorporate an automated decision making process are obliged to provide customers who ask a reason and explanation for refusing a loan, a job application etc.