INDIAN INSTITUTE OF TECHNOLOGY
JODHPUR

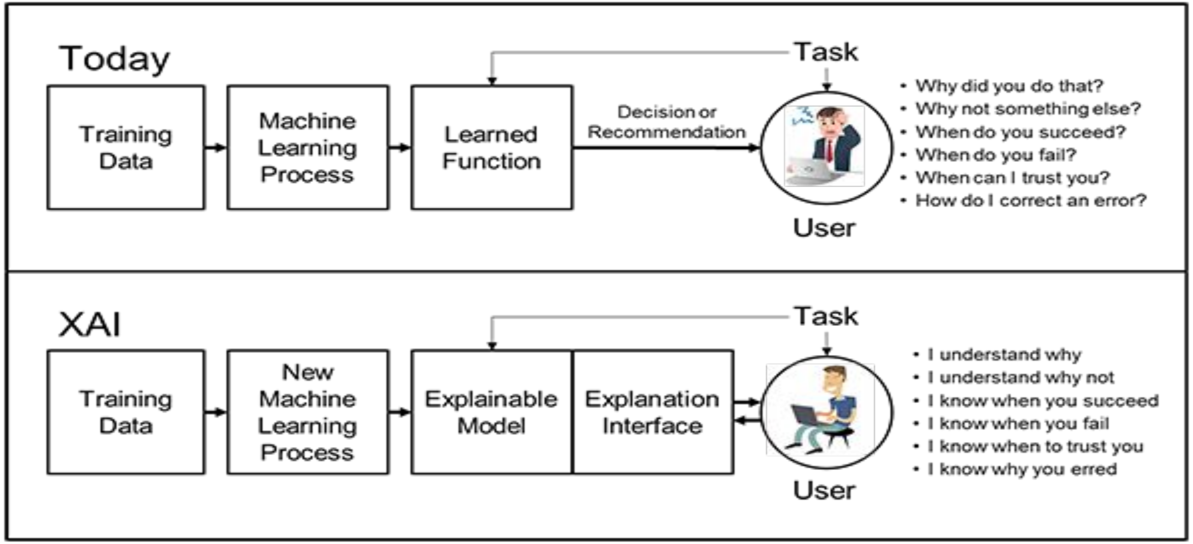
DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Explainable AI Techniques

A presentation by Sagar Kumar Agrawal , Dinesh Kumar and Devesh Nain

Explainable AI Concept

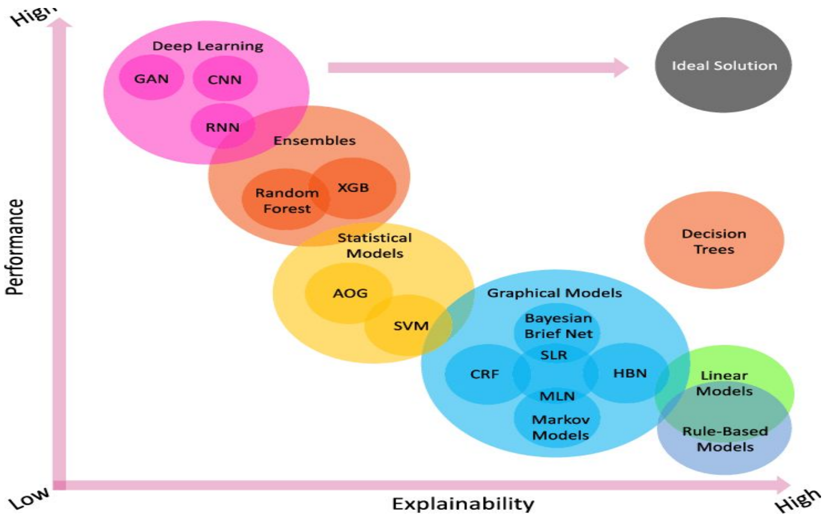
DARPA XAI



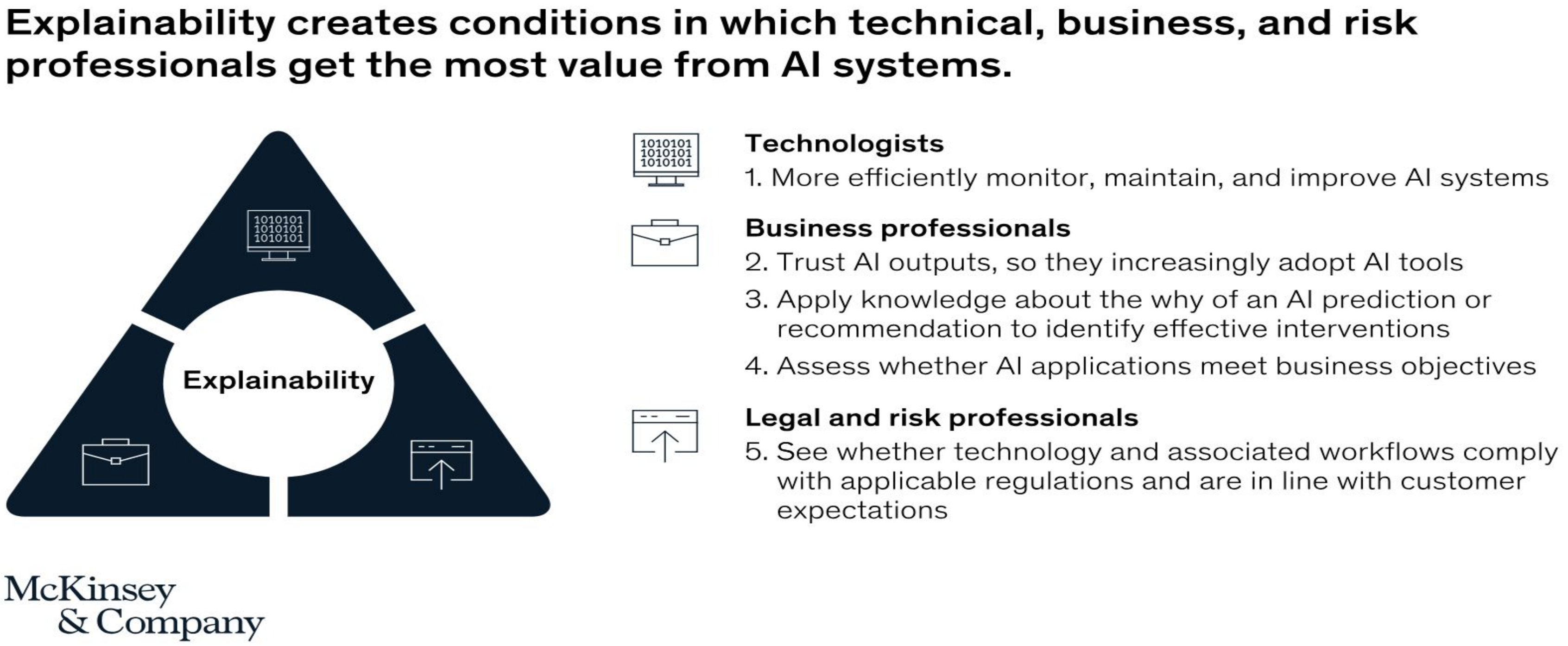
Accuracy Vs Explainability

Deep Learning Models - High Accuracy Low Explainability

Linear / Tree based Models - Low Accuracy High Explainability



Why-businesses-need-explainable-AI?



Explainability creates conditions in which technical, business, and risk professionals get the most value from AI systems.

Technologists

1. More efficiently monitor, maintain, and improve AI systems

Business professionals

2. Trust AI outputs, so they increasingly adopt AI tools
3. Apply knowledge about the why of an AI prediction or recommendation to identify effective interventions
4. Assess whether AI applications meet business objectives

Legal and risk professionals

5. See whether technology and associated workflows comply with applicable regulations and are in line with customer expectations

McKinsey & Company

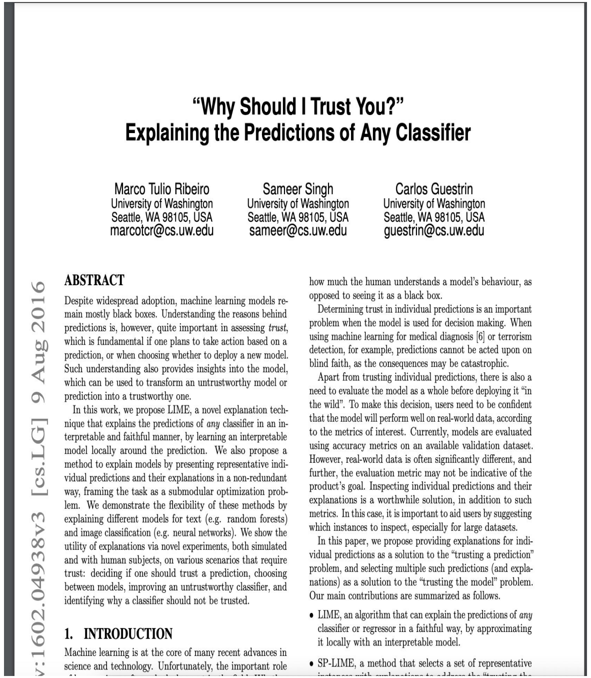
LIME : Local Interpretable Model-Agnostic Explanations

LIME Was Introduced in this 2016 paper **“Why Should I Trust You?” Explaining the Predictions of Any Classifier**.

It aims to explain any black box model using local approximations and has been one of the most cited paper in field of Explainable Artificial Intelligence

LIME focuses on training local surrogate models to explain individual predictions.

Local surrogate models are interpretable models that are used to explain individual predictions of black box machine learning models.



“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

v:1602.04938v3 [cs.LG] 9 Aug 2016

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role

how much the human understands a model’s behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product’s goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- SP-LIME, a method that selects a set of representative

LIME : Intuition

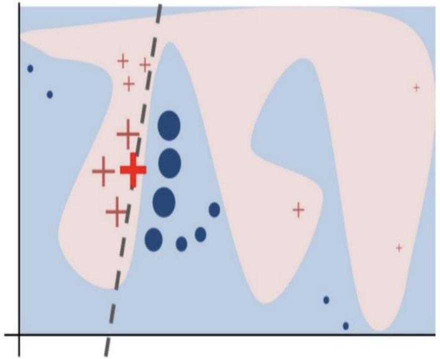


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

The way of interpretation is to perturb this sample (thick + sign) to obtain other points on the graph, and these points can obtain predicted values (marked + or **O**) in the original model.

Then, fit a simple model (simple linear regression) to these samples, which is the gray dotted line in the figure, and then use this simple model to explain which features affect the output results for the "thick +" sample.

From a global perspective, the performance of this linear model is much worse than that of the complex model (the simple linear model will only classify the left half as Negative and the right half as Positive), but in the vicinity of the "thick +" sign, this The performance of the simple model is actually not bad.

[Ribeiro et al 2016]

LIME

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

- The explanation model for instance x is the model g (e.g. linear regression model) that minimizes loss L (e.g. mean squared error), which measures how close the explanation is to the prediction of the original model f (e.g. an xgboost model). (**How close XGBOOST is to Linear regression model in local region?**)
- G is the family of possible explanations, for example all possible linear regression models.
- The proximity measure πx defines how large the neighborhood around instance x is that we consider for the explanation.
- In practice, LIME only optimizes the loss part.
- The user has to determine the complexity, e.g. by selecting the maximum number of features that the linear regression model may use.
- While the model complexity $\Omega(g)$ is kept low (e.g. prefer fewer features). Governed by users

LIME : Training Surrogate Models

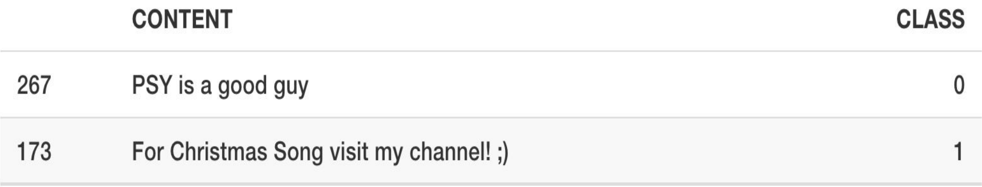
The recipe for training local surrogate models:

- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new points.
- LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model.
- On this new dataset LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest.
- Explain the prediction by interpreting the local model.

Reference : https://christophm.github.io/interpretable-ml-book/lime.html

LIME : Text Data

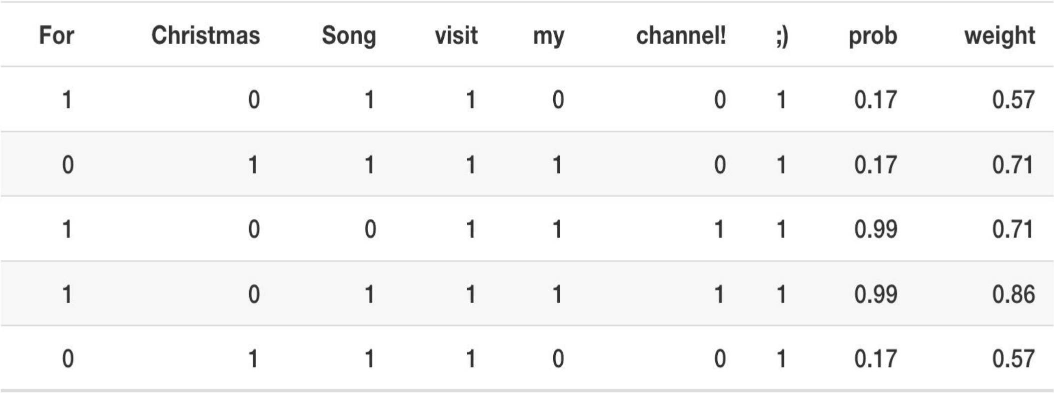
- In this example we classify YouTube comments as spam or normal.
- The black box model is a deep neural network trained on the document word matrix.
- Let us look at the two comments of this dataset and the corresponding classes (1 for spam, 0 for normal comment):



	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channel! ;)	1

LIME : Text Data

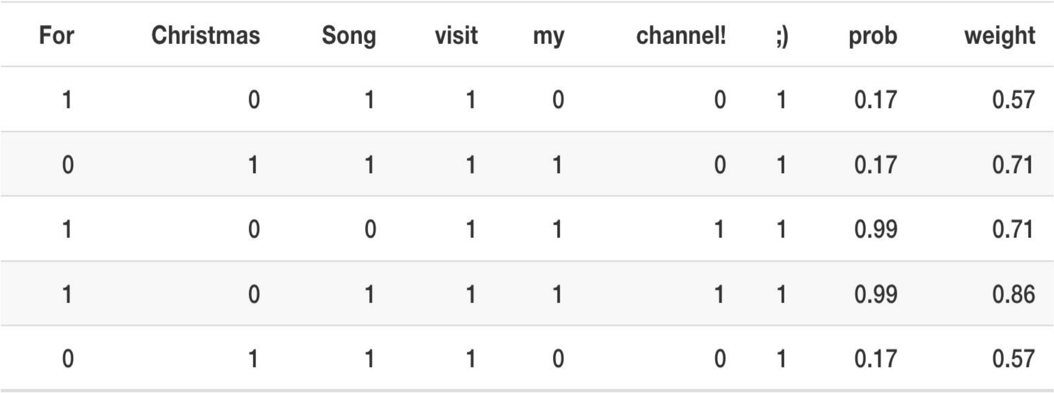
- The next step is to create some variations of the datasets used in a local model.
- Each row is a variation, 1 means that the word is part of this variation and 0 means that the word has been removed.
- The “prob” column shows the predicted probability of spam for each of the sentence variations.
- The “weight” column shows the proximity of the variation to the original sentence, calculated as 1 minus the proportion of words that were removed,



For	Christmas	Song	visit	my	channel!	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

LIME : Text Data

- Here are the two sentences (one spam, one no spam) with their estimated local weights found by the LIME algorithm:
- The word “channel” indicates a high probability of spam



For	Christmas	Song	visit	my	channel!	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57