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

Authors:

Marco Tulio Ribeiro
Sameer Singh
Carlos Guestrin
(University Of Washington)

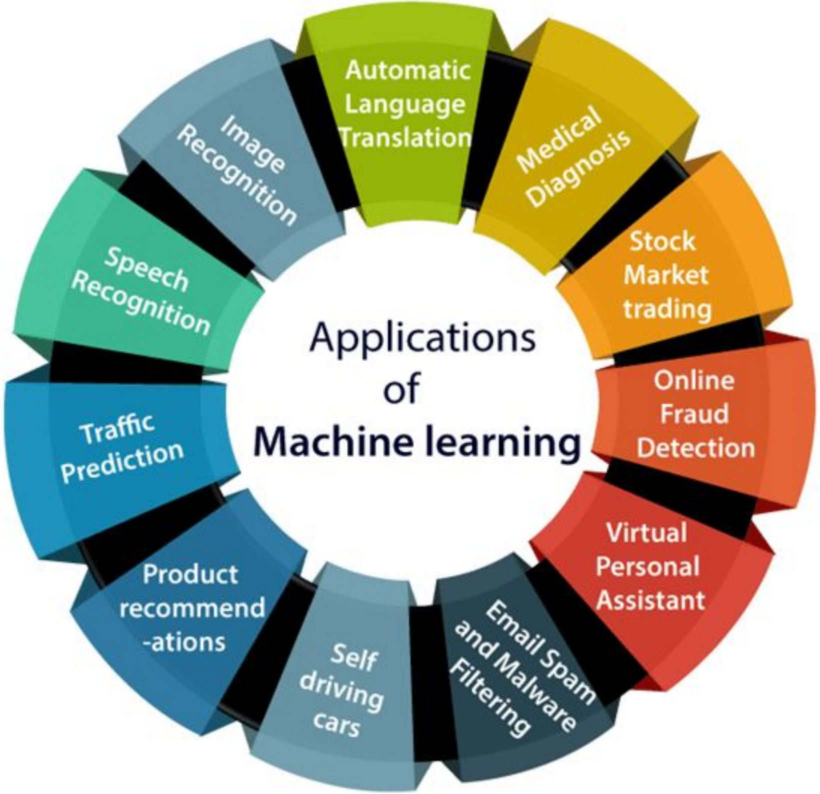
Conference :

KKD-2016
22nd Conference on Knowledge discovery and Data mining

Presentation :

Nooshin SHOJAEE
STAI_AI Ethics
0190592333

Why do we need an explainer for a Machine learning model ?



Machine learning is at the core of science and technology

Why do we need an explainer for a Machine learning model ?

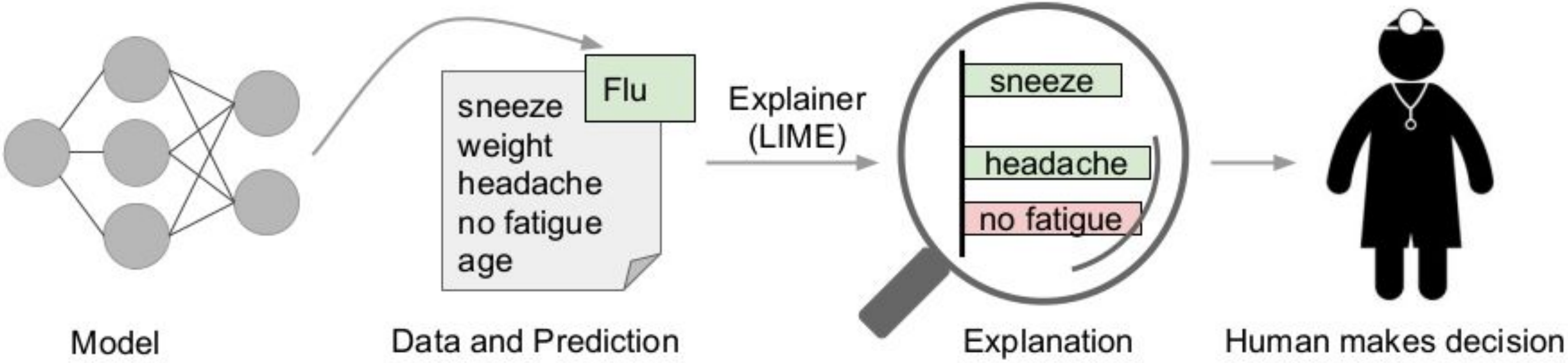
Data → Machine learning model → Prediction and Decision → **How do you Trust this ??**

Trusting a prediction

Trusting a model

Deploy a model

The case for explanations



The case for explanations

Common but not efficient solutions :

Interpretable models

Decision trees

Accuracy - Interpretability
Trade off

Measuring Accuracy

Cross validation

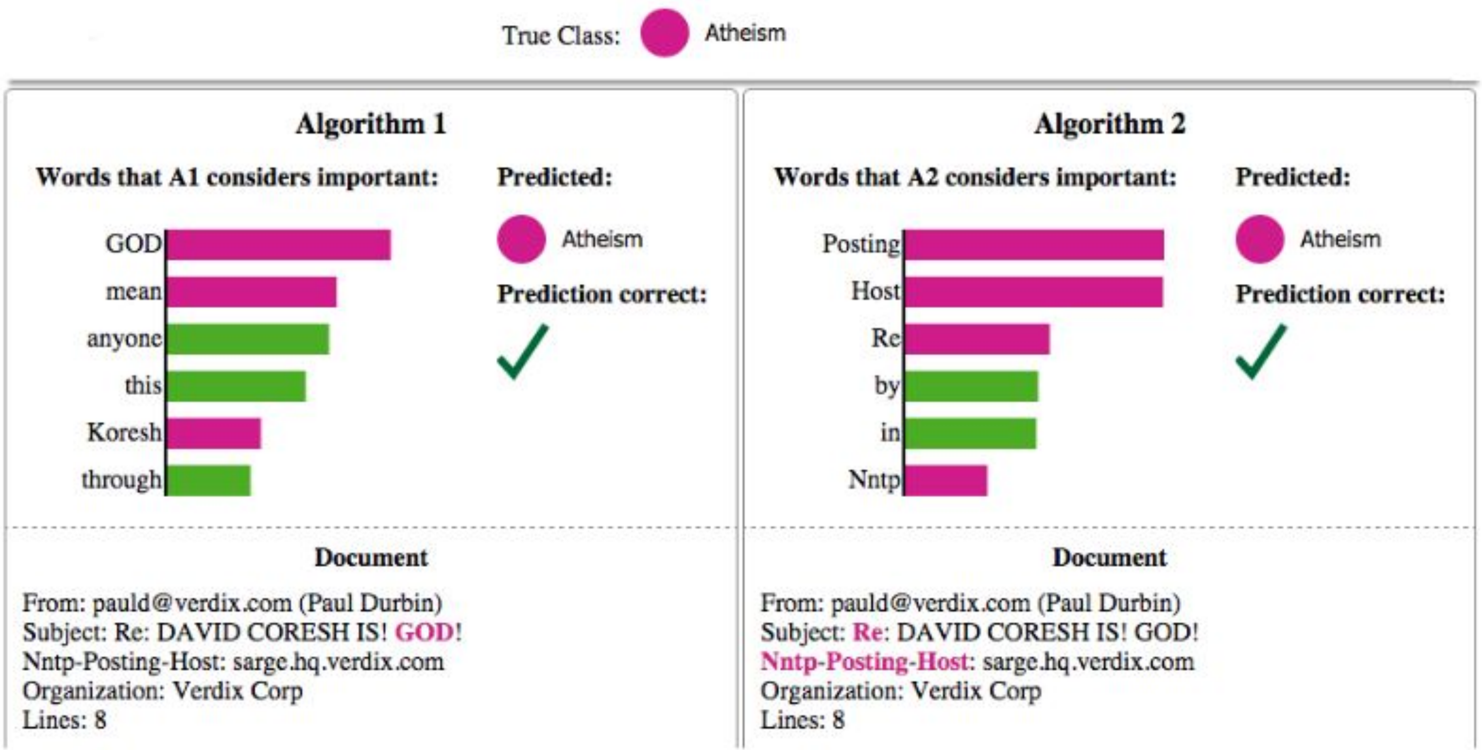
Data leakage
(Fake accuracy)

Test on real world datas

A/B testing

Very expensive

The case for explanations



The case for explanations

Desired Characteristics for Explainers

Interpretable

Local fidelity

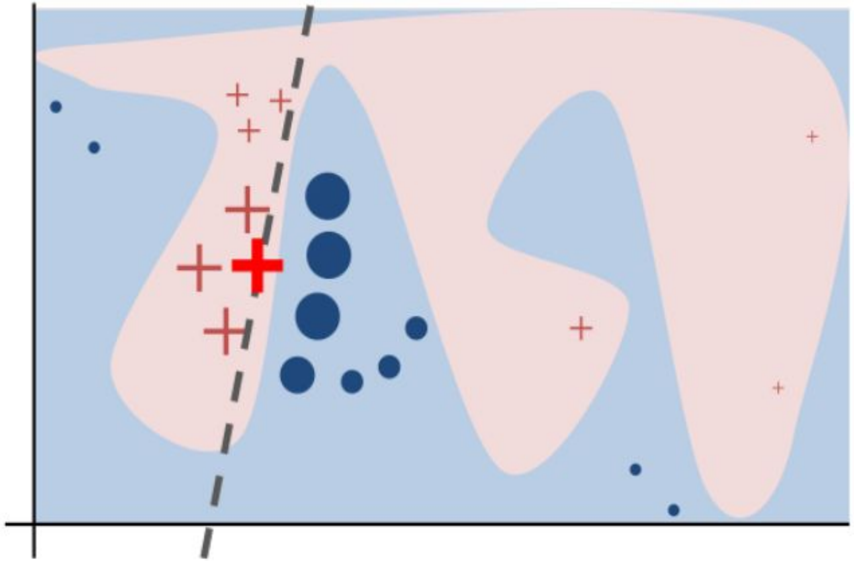
model-agnostic

Global perspective

LIME : **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$



X: explained sample
F: black box classifier
G: explanation model
Z: perturbed sample
L: unsimilarity between g and f
Π:locality around x

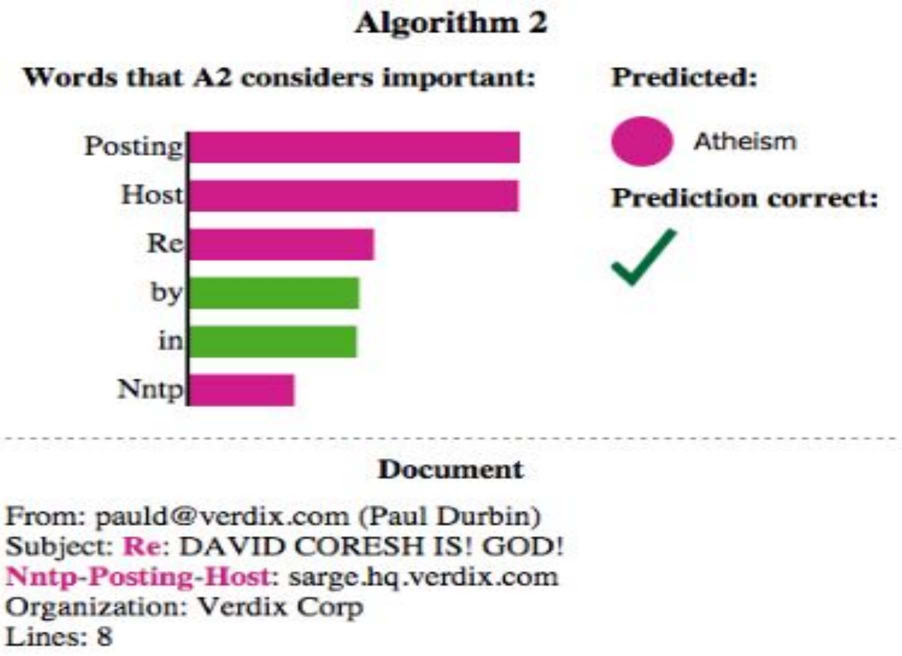
LIME : **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

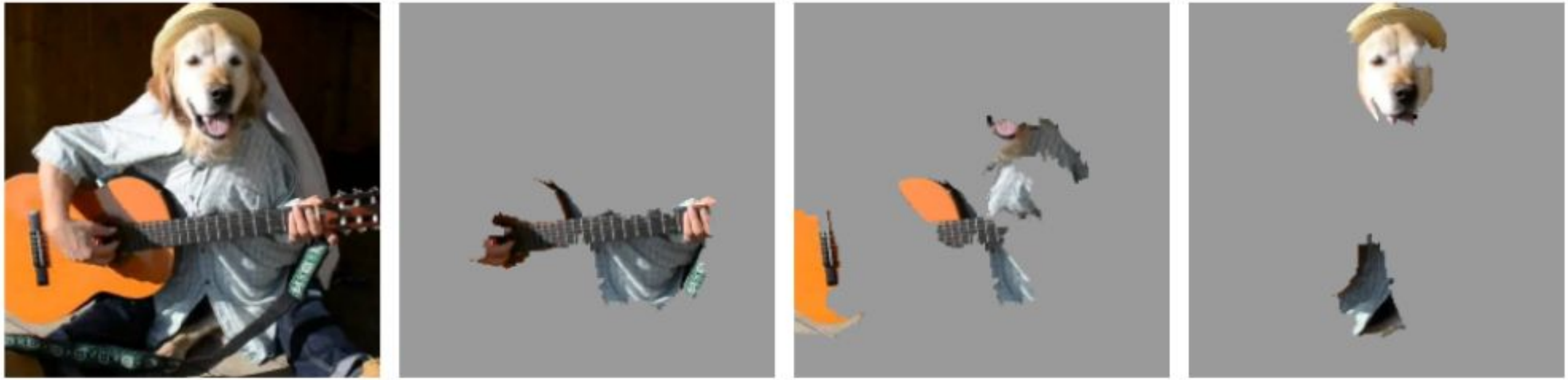
Example 1 : Text classification with SVMs



20 newsgroup data set

Accuracy :94%

Example 2 : Deep networks for images



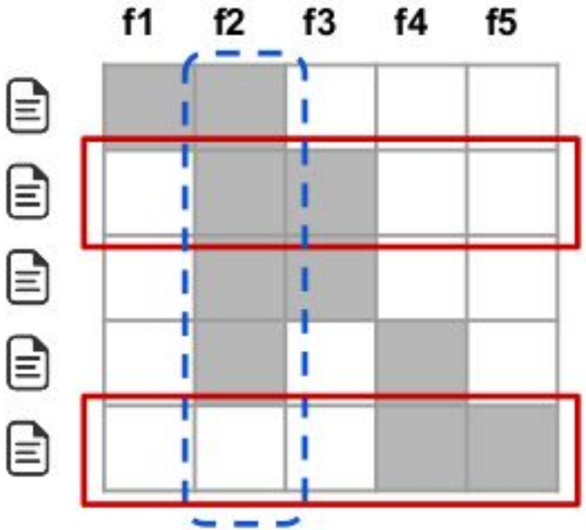
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Classification prediction by Google’s Inception neural network
Electric Guitar (p = 0.32)
Acoustic guitar (p = 0.24)
Labrador (p = 0.21)

SP-LIME : Submodular Pick for explanation models

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j$$

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}}\, c(V, \mathcal{W}, I)$$



C: coverage (total importance of the feature)
V: set of explanations
W: Matrice of instance-feature
I: global feature importance

SP-LIME : Submodular Pick for explanation models

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B

  for all $x_i \in X$ **do**

    $\mathcal{W}_i \leftarrow \textbf{explain}(x_i, x'_i)$ ▷ Using Algorithm $\boxed{1}$

  end for

  for $j \in \{1 \dots d'\}$ **do**

    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$ ▷ Compute feature importances

  end for

  $V \leftarrow \{\}$

  while $|V| < B$ **do** ▷ Greedy optimization of Eq ($\boxed{4}$)

    $V \leftarrow V \cup \text{argmax}_i \, c(V \cup \{i\}, \mathcal{W}, I)$

  end while

  return V

Simulated User Experiments

Experiment setup

Data set : reviews on Books and DVDs (2000 instances each)

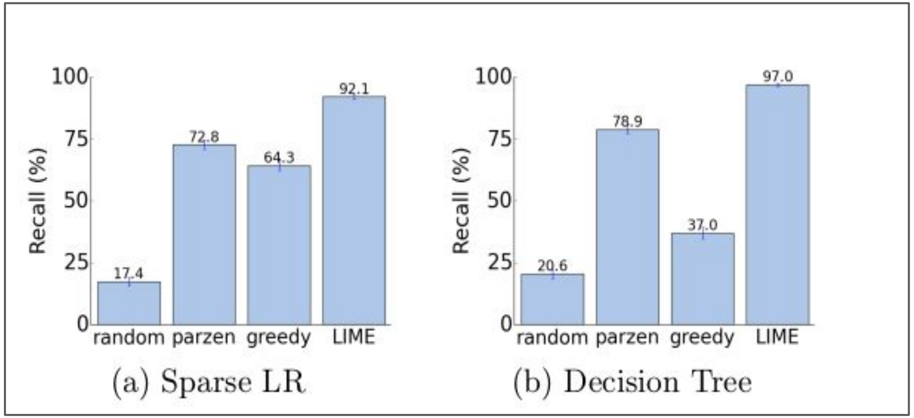
Classification Problem : positive and negative reviews

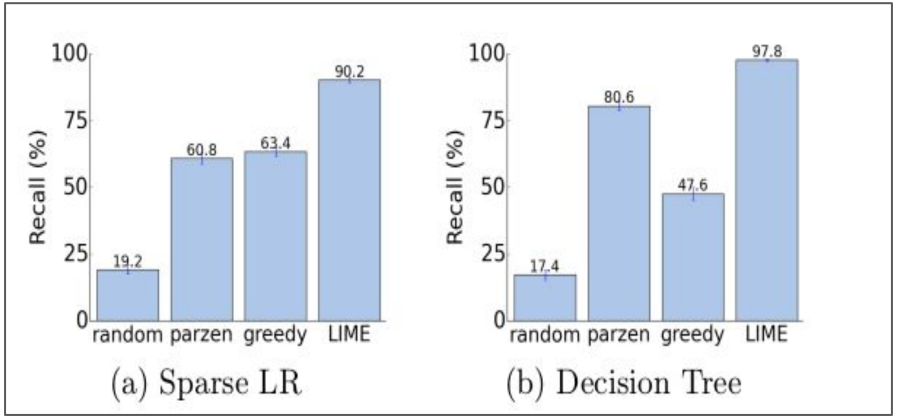
Classification models : DT,NN,LR,SVM,RF

Explainers : LIME,Parzen,greedy,Random

Simulated User Experiments

Are the explanations faithful to the model?





Books

DVDs

Recall on truly important features for two interpretable classifiers on the Books/DVDs dataset

Simulated User Experiments

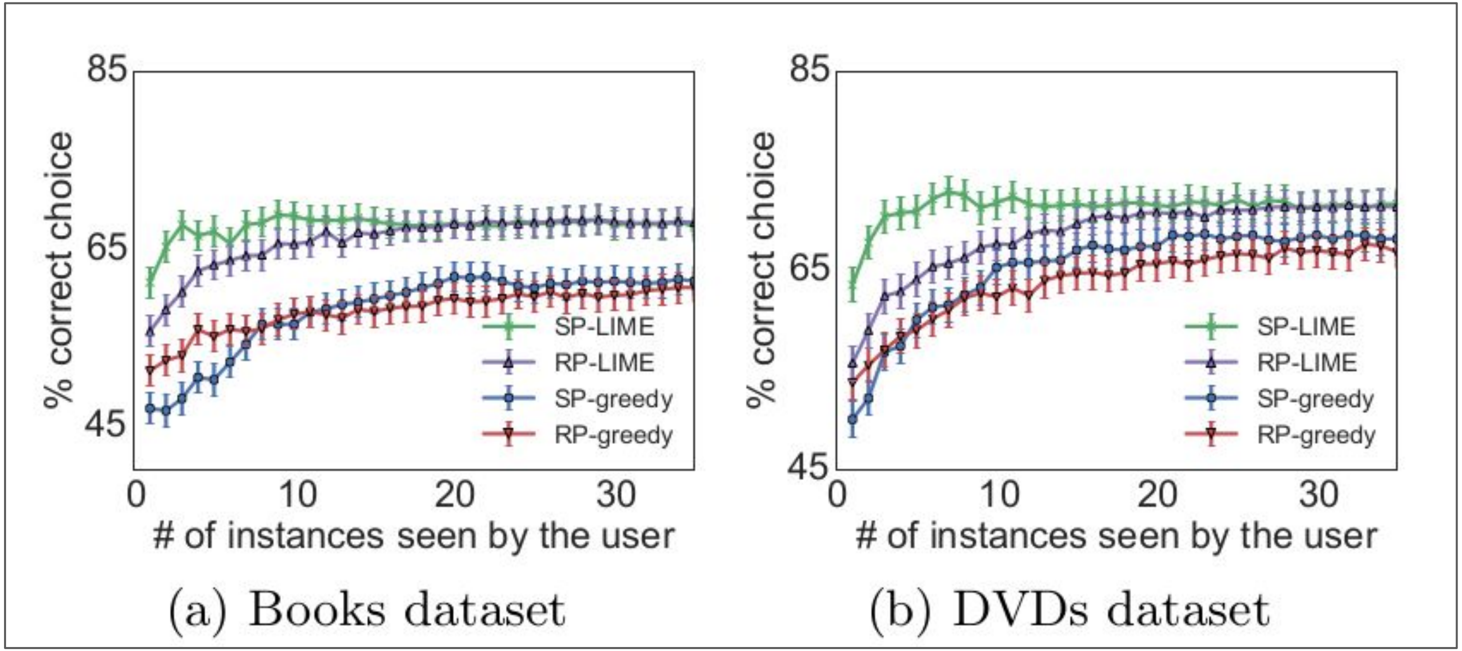
Should I trust this prediction ?

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

Average F1 of trustworthiness for different explainers on a collection of classifiers and datasets

Simulated User Experiments

Can I trust this model?



Choosing between two classifiers, as the number of instances shown to a simulated user is varied.

Evaluation With human subjects

Experiment setup

Training Data set 1: 20 newsgroup
Training Data set 2 : *Cleaned* 20 newsgroup

Test Data set : 20 news group
Test Data set : Religion data set

Classification Problem : Christianity vs. Atheism

Classification models : SVM , *cleaned* SVM

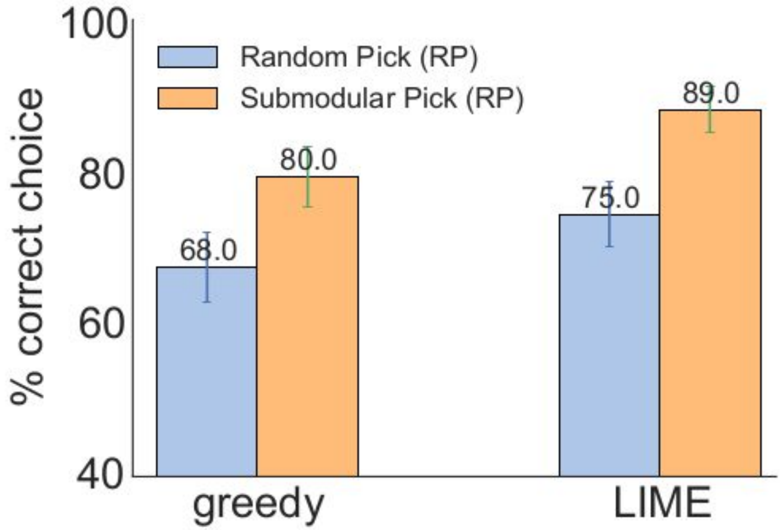
Explainers : LIME,greedy

Evaluation With human subjects

Can users select the best classifiers ?

	SVM	Cleaned SVM
Religion	57.3%	69.0%
20 newsgroup	94%	88.6%

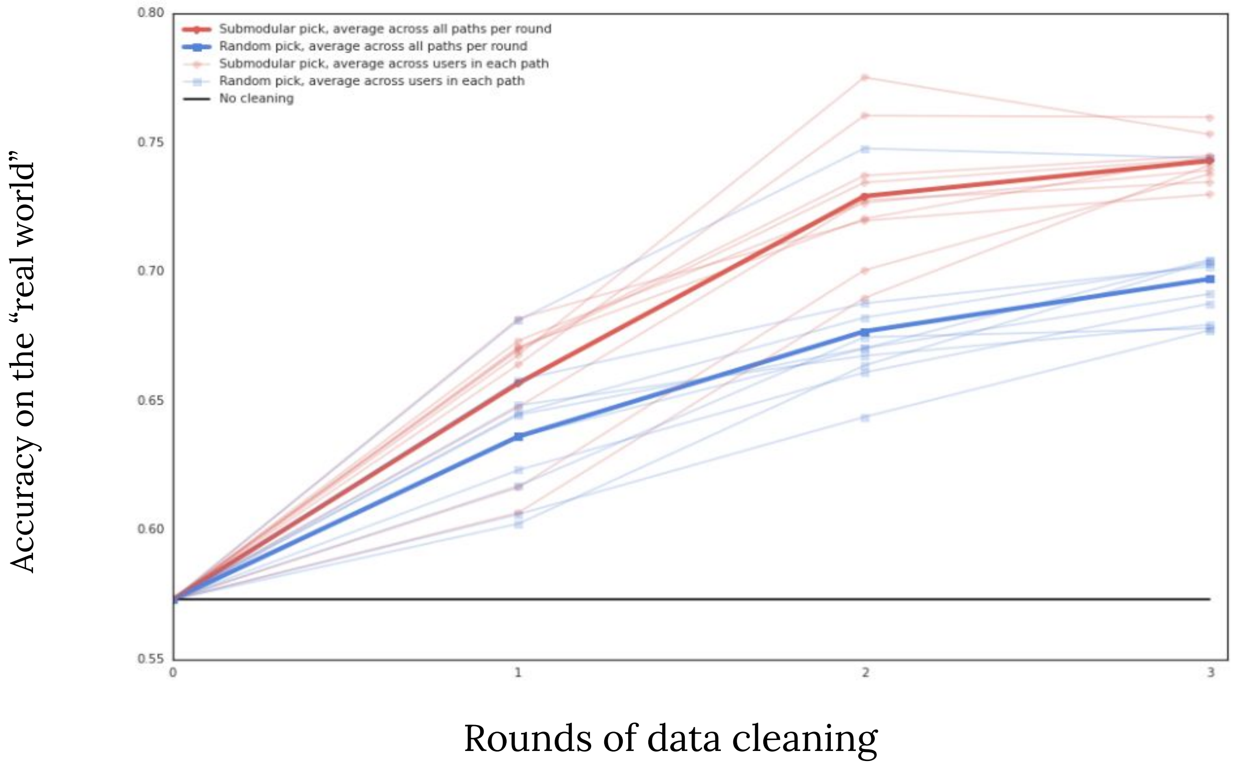
Accuracy measure



Average accuracy of human subject in choosing between two classifiers

Evaluation With human subjects

Can non experts improve the classifier ?



- Submodular pick LIME
- Random pick LIME

Evaluation With human subjects

Do Explanations lead to insight ?

(a) Husky classified as wolf (b) Explanation

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

"Husky vs Wolf" experiment results

Related works

- Gelsat
- Modeltracker
- Letting user know when the systems are likely to to fail
- Exposing users to different kind of mistakes
- Using interpretable models in medical domain
- Eluci debug for text
- Computer vision systems (alignment)
- Gradient vector as explanation

Conclusion and Future works

- Importance of trust in human-Machine learning systems interactions
- Potential of explainability in assessing trust
- Proposing LIME as an approach to explain the prediction of any model
- Introducing SP-LIME providing a global view of any model
- With explainability even non experts can achieve feature engineering

- Fix pick step issue when Decision tree is used as the explanation model
- Investigate in other domains : speech,video,medical ,etc.

Google Scholar

Marco Tulio Ribeiro

FOLLOW

Microsoft Research
Verified email at cs.washington.edu - Homepage

Machine Learning   Natural Language Processing

TITLE	CITED BY	YEAR
" Why Should I Trust You?": Explaining the Predictions of Any Classifier MT Ribeiro, S Singh, C Guestrin Knowledge Discovery and Data Mining (ACM KDD)	4193	2016
Anchors: High-Precision Model-Agnostic Explanations MT Ribeiro, S Singh, C Guestrin AAAI	471	2018
Model-agnostic interpretability of machine learning MT Ribeiro, S Singh, C Guestrin arXiv preprint arXiv:1606.05386	250	2016

LIME & GDPR

References

S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard,and J. Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In Human Factors in Computing Systems (CHI), 2015.

D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual clas-sification decisions. Journal of Machine Learning Research,11, 2010.

K. Patel, J. Fogarty, J. A. Landay, and B. Harrison. Inves-tigating statistical machine learning as a tool for software development. In Human Factors in Computing Systems (CHI), 2008.

https://www.youtube.com/watch?v=KP7-JtFMLo4&t=932s

https://medium.com/@thommash/local-interpretable-model-agnostic-explanations-lime-and-gdpr-9e3d66b64207

Thank you for your attention :)