

LIME - Local Interpretable Model-agnostic Explanations

Tulio Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016).
“Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97-101. 10.18653/v1/N16-3020.







Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

Tylko 1 błąd!



Predicted: **wolf**
True: **husky**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **husky**

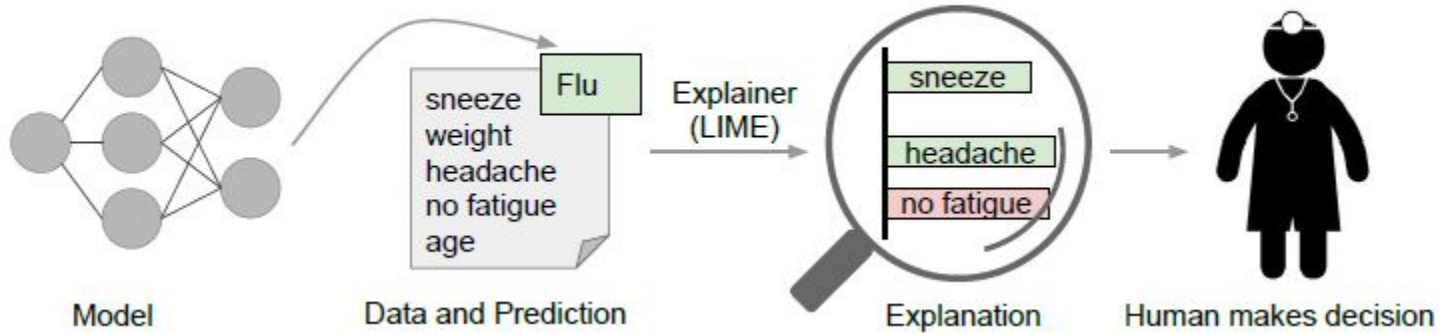


Predicted: **husky**
True: **husky**

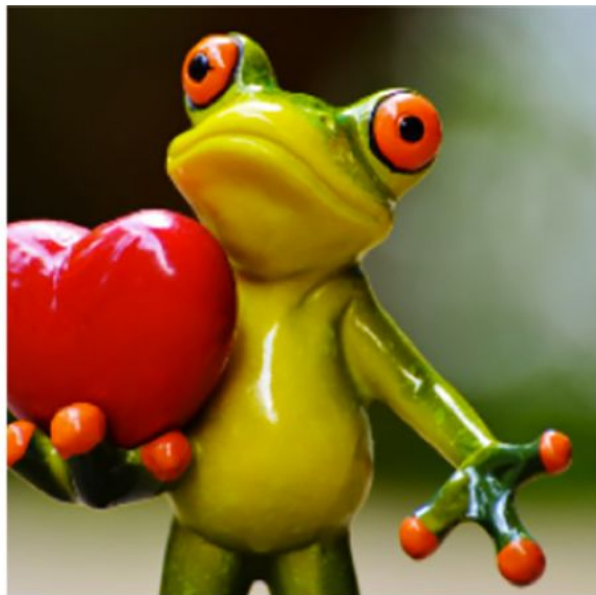


Predicted: **wolf**
True: **wolf**

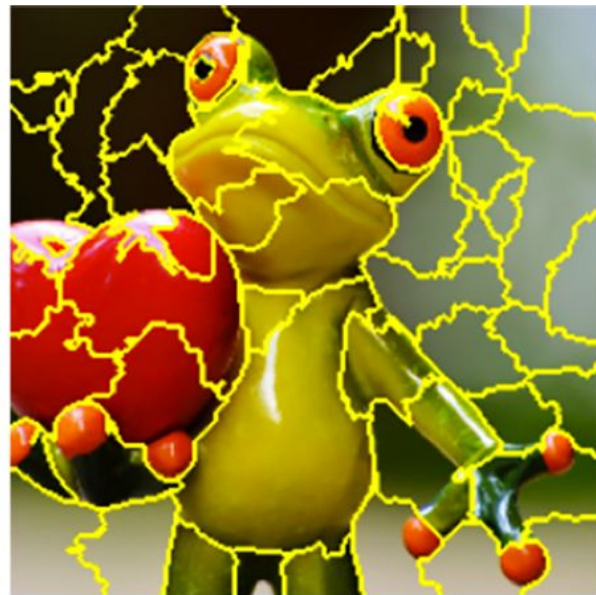
Wyjaśnianie pojedynczej predykcji



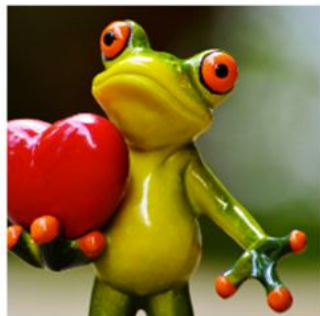
Jak działa LIME? - przypadek szczególny



Original Image



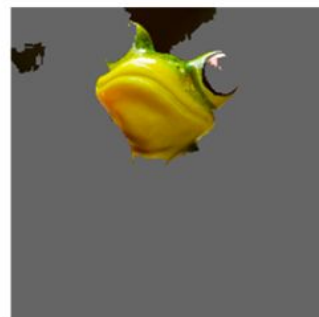
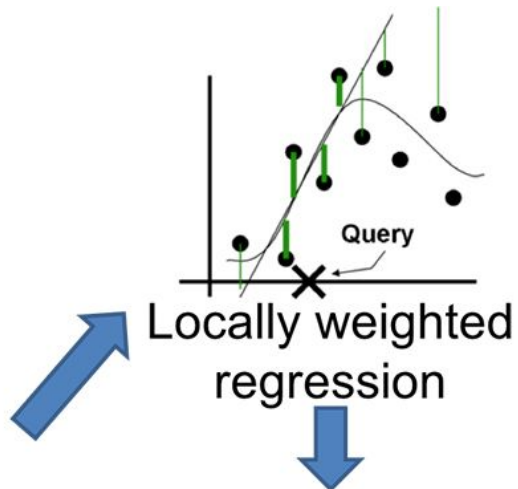
Interpretable
Components



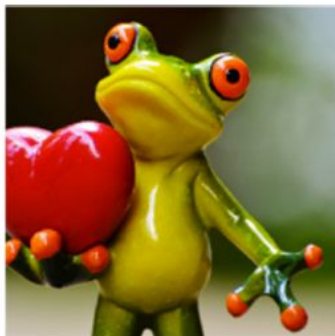
Original Image
 $P(\text{tree frog}) = 0.54$



| Perturbed Instances | $P(\text{tree frog})$ |
|---------------------|-----------------------|
| | 0.85 |
| | 0.00001 |
| | 0.52 |



Explanation



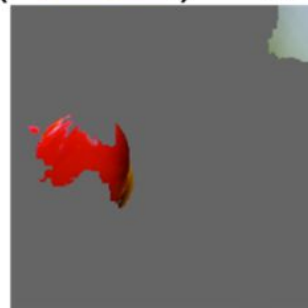
$$P(\text{img}) = 0.54$$



$$P(\text{img}) = 0.07$$



$$P(\text{img}) = 0.05$$



Który model wybrać?

Example #3 of 6

True Class:  Atheism

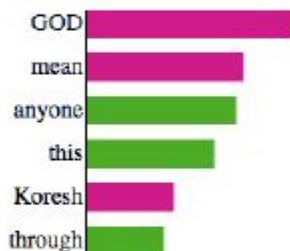
[Instructions](#)

[Previous](#)

[Next](#)

Algorithm 1

Words that A1 considers important:



Predicted:

 Atheism

Prediction correct:



Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD!**
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:



Predicted:

 Atheism

Prediction correct:



Document

From: pauld@verdix.com (Paul Durbin)
Subject: **Re:** DAVID CORESH IS! **GOD!**
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Pożądane własności



interpretowalność



lokalna wierność



niezależność od modelu

A można matematycznie?

$x \in \mathbb{R}^d$ - oryginalna reprezentacja rozpatrywanego przypadku (obserwacji)

$x' \in 0, 1^{d'}$ - wektor binarny, interpretowalna reprezentacja x

$g \in G$ - wyjaśniający model z klasy potencjalnie interpretowalnych modeli, np modele liniowe, drzewa, ...

$\Omega(g)$ - miara złożoności modelu

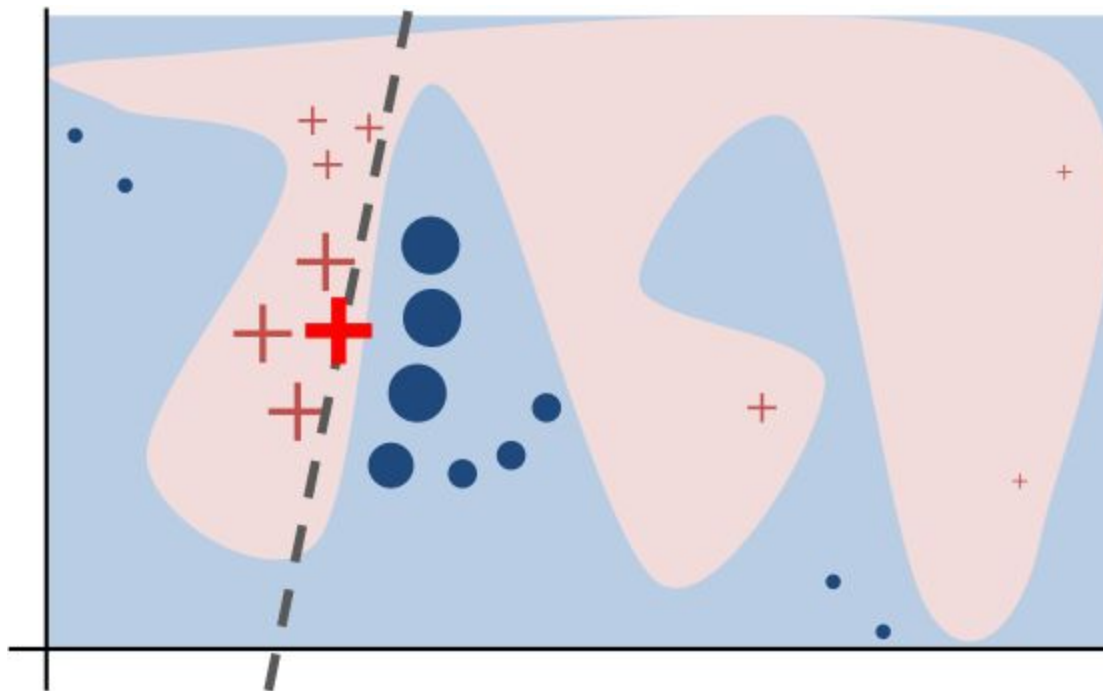
$f : \mathbb{R}^d \rightarrow \mathbb{R}$ - model wyjaśniany

$\pi_x(z)$ - miara odległości obserwacji z od x

$\mathcal{L}(f, g, \pi_x)$ - miara wiarygodności modelu g w przybliżaniu modelu f .

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

Dopasowanie lokalnego modelu



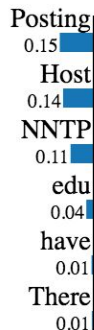
Lokalne wyjaśnienie modelu

Prediction probabilities



atheism

christian



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.

R - use case

```
> library(lime)
> explainer <- lime(sentence_to_explain, model = xgb_model, preprocess = get_matrix)
> explanation <- explain(sentence_to_explain, explainer, n_labels = 1, n_features = 2)
```

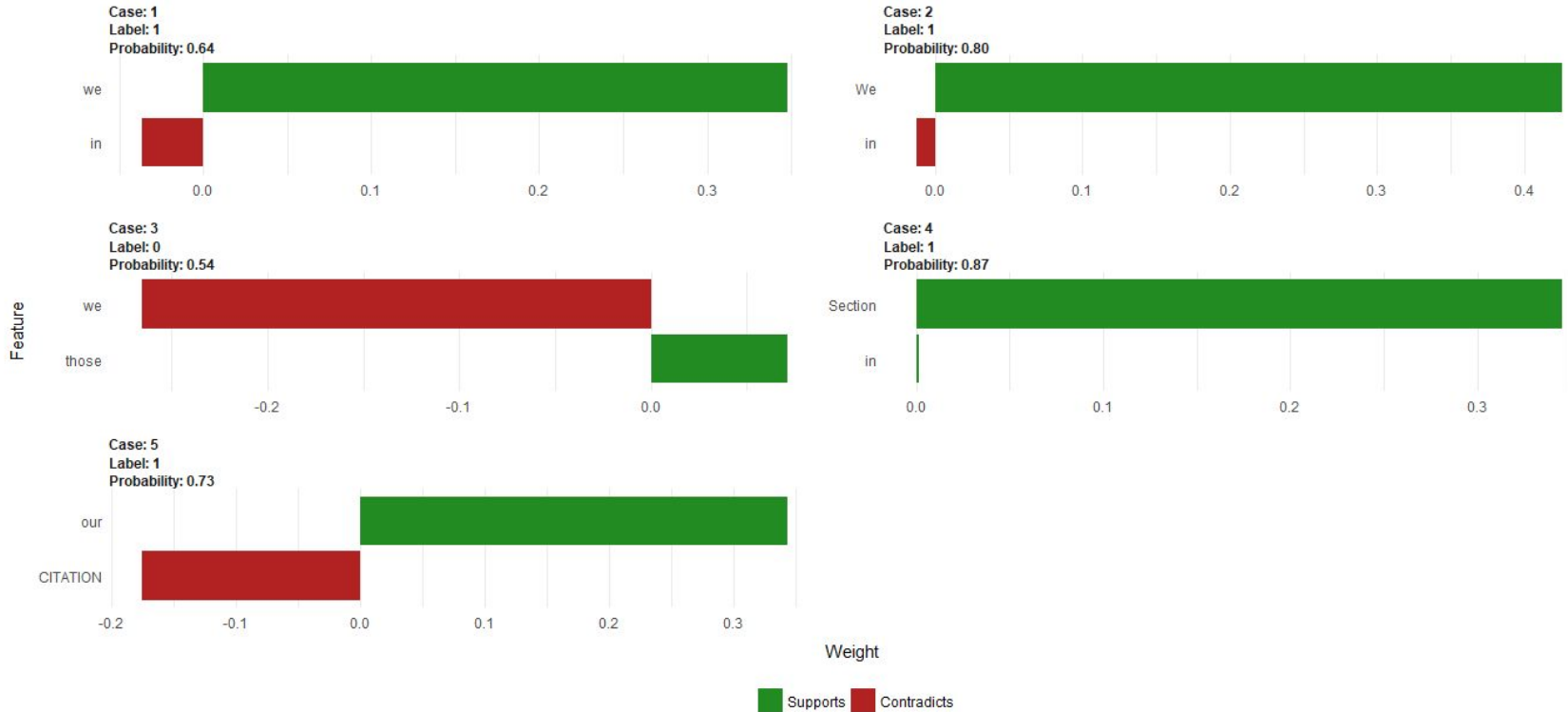
R - use case

```
> library(lime)
> explainer <- lime(sentence_to_explain, model = xgb_model, preprocess = get_matrix)
> explanation <- explain(sentence_to_explain, explainer, n_labels = 1, n_features = 2)
> explanation[, 2:9]
```

| | case | label | label_prob | model_r2 | model_intercept | model_prediction | feature | feature_value |
|----------|------|-------|------------|-----------|-----------------|------------------|----------|---------------|
| in | 1 | 1 | 0.6418385 | 0.9881959 | 0.3323441 | 0.6436369 | in | in |
| we | 1 | 1 | 0.6418385 | 0.9881959 | 0.3323441 | 0.6436369 | we | we |
| in1 | 2 | 1 | 0.8022363 | 0.8535779 | 0.3223901 | 0.7353984 | in | in |
| We | 2 | 1 | 0.8022363 | 0.8535779 | 0.3223901 | 0.7353984 | We | We |
| those | 3 | 0 | 0.5432571 | 0.8854101 | 0.6918100 | 0.4974141 | those | those |
| we1 | 3 | 0 | 0.5432571 | 0.8854101 | 0.6918100 | 0.4974141 | we | we |
| in2 | 4 | 1 | 0.8719526 | 0.6674597 | 0.4926221 | 0.8397768 | in | in |
| Section | 4 | 1 | 0.8719526 | 0.6674597 | 0.4926221 | 0.8397768 | Section | Section |
| CITATION | 5 | 1 | 0.7316587 | 0.4420266 | 0.4437610 | 0.6119218 | CITATION | CITATION |
| our | 5 | 1 | 0.7316587 | 0.4420266 | 0.4437610 | 0.6119218 | our | our |

R - use case

```
> plot_features(explanation)
```



Local Interpretable Model-agnostic Explanations

Put here the text to explain

I

Quantity of permutations to generate

5000

Word selection strategies

auto

Number of words to select

1 2 20

Text provided is too short to be explained (≥ 3).

*There once was a package called lime,
Whose models were simply sublime,
It gave explanations for their variations,
one observation at a time.*



lime-rick by Mara Averick

Kilka Źródeł na koniec

<https://arxiv.org/abs/1602.04938>

https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html

<https://www.slideshare.net/0xdata/explaining-blackbox-machine-learning-predictions>

<https://www.youtube.com/watch?v=KP7-JtFMLo4>

<https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime>

<https://github.com/marcotcr/lime>