“Why Should You Trust My Explanation?”
Understanding Uncertainty in LIME Explanations

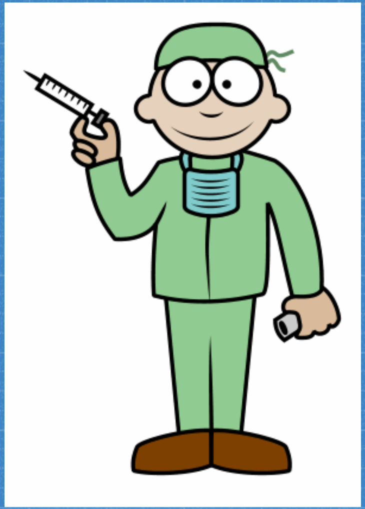
Outline

- Why Lime?
- What is LIME?
- How Lime works?
- Uncertainty in Lime

Sometimes you don’t know if you can trust a machine learning prediction..

Janet has flu

Why should I trust you?



Its easier to trust a prediction if you understand the reasons for it..

Janet has flu

Symptoms
- Fever
- Headache
- Fatigue

Okay, I trust you now.

Or to figure out when you shouldn’t trust a model..

I am 100% sure this is a wolf

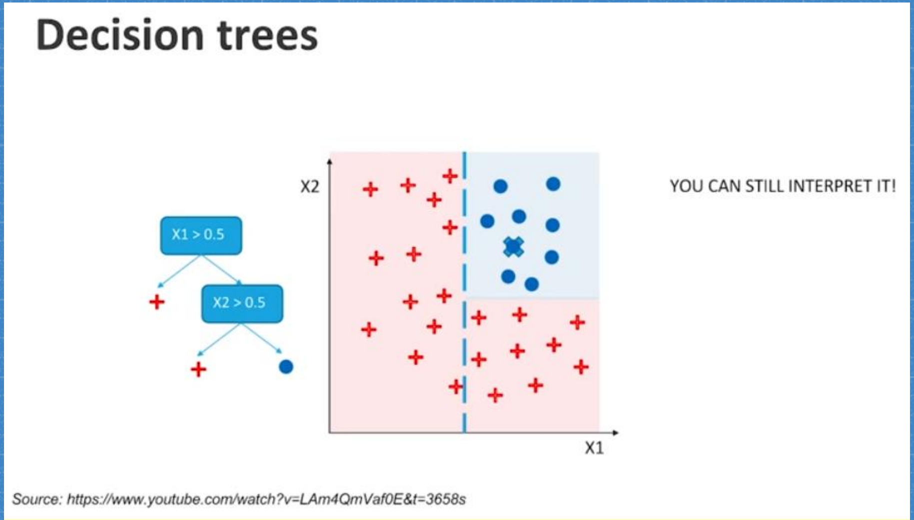
Hmm..

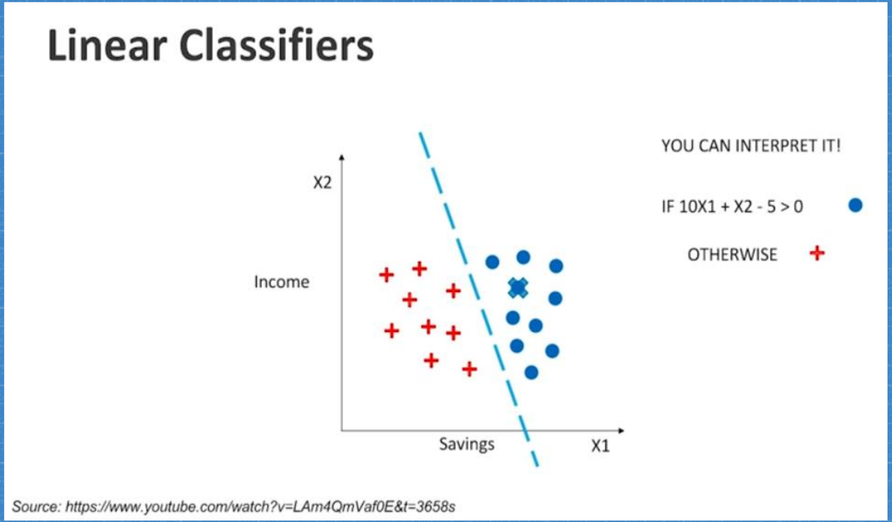
Or to figure out when you shouldn’t trust a model..

I am 100% sure
this is a wolf

You are detecting
snow, not wolves!
I don’t trust you!

Examples of Interpretable models

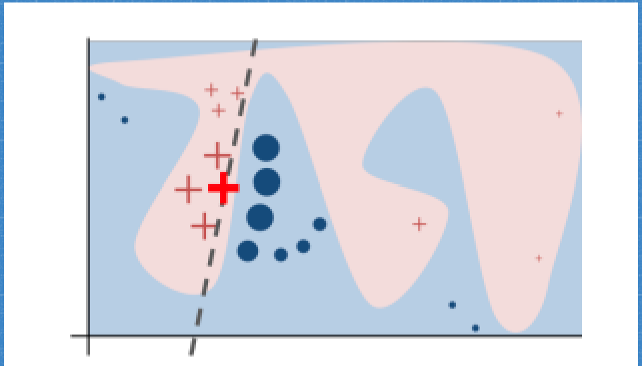




What is a black box model?

Input → BLACK BOX → Output

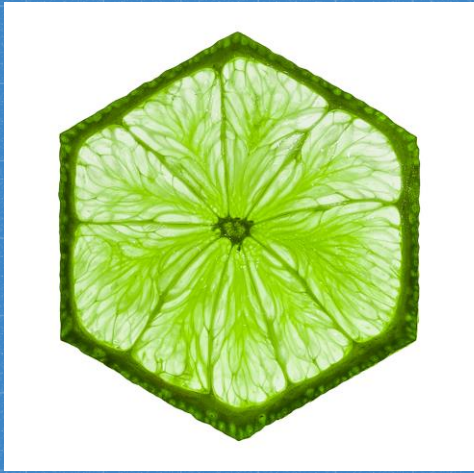
A system where the internal workings are completely hidden from you.
Eg: Deep Neural Network



Source : Ribeiro et al, 2016

What if you could understand why any model is making a prediction..

The Lime Algorithm



GOAL: Understand the prediction of an arbitrary model for a certain sample.

LIME : Local Interpretable Model-agnostic Explanations

Local

Explanations are locally faithful instead of globally

Interpretable

Humans are limited by an amount of information that can be processed and understood.

E.g., the weights of a neural network are not meaningful for a human.

Model-Agnostic

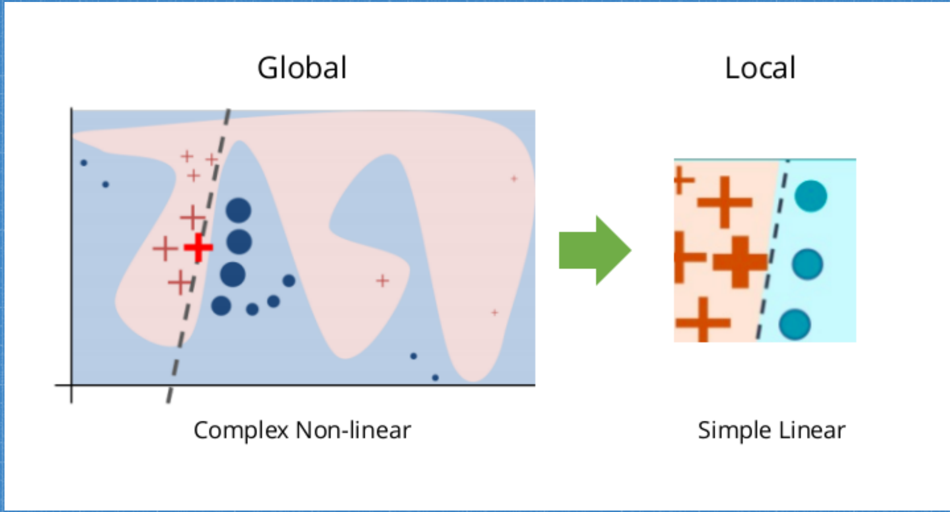
Any machine learning algorithm can be used as predictive model. Works with text, image and tabular data.

Explanations

Artifacts that explain the relationship between a model's input and its prediction.

How it works?

- Generate a fake dataset X *from the example.*
- Use trained black-box model f to get predictions y_p for each example in the generated dataset.
- Train a white-box model g on X, y_p.
- Explain the original example through weights of the white-box model.
- Assess how well the white-box model approximates the black-box model.



Source : Ribeiro et al, 2016

The math in Lime

$$\xi(x) = \text{argmin}_{g \in G} \overbrace{\mathcal{L}(f, g, \Pi_x)}^{\text{Good approximation}} + \overbrace{\Omega(g)}^{\text{Stay simple}}$$

Explanation

Family of interpretable models

Complex model

Simple interpretable model

Proximity

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

Kernel distance of z from x

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

Model Label

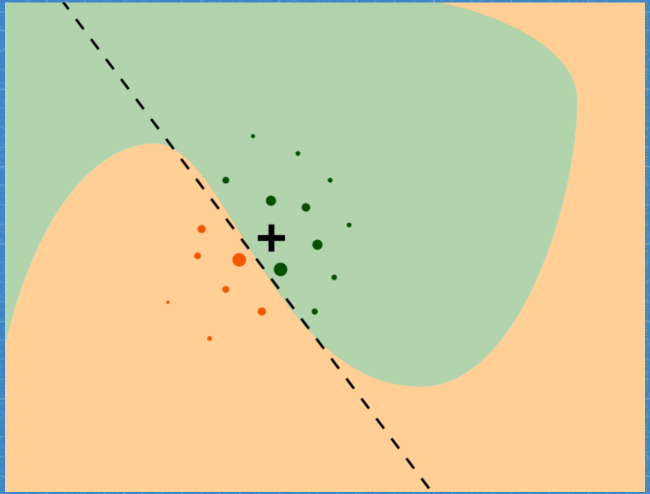
Interpretable model prediction

$$\Pi_x(z) = \exp(-D(x, z)^2 / \sigma^2)$$

New Dataset

Labels: Prediction of complex model

Features: newly generated datapoints



Source: https://www.pdfprof.com/PDF_Image.php?idt=31649&t=27

Example: Text based Classifier

Interpretable input

$$z'_i \leftarrow \text{sample\_around}(x_i)$$

Model : deep decision tree trained on the document word matrix

For	Christmas	Song	Visit	My	Channel!	;)	P(Spam)	Weight
✓	✗	✓	✓	✗	✗	✓	0.17	0.57
✗	✓	✓	✓	✓	✗	✓	0.17	0.71
✓	✗	✗	✓	✓	✓	✓	0.99	0.71
✓	✗	✓	✓	✓	✓	✓	0.99	0.86
✗	✓	✓	✓	✗	✗	✓	0.17	0.57

Source: https://christophm.github.io/interpretable-ml-book/lime.html#lime-for-text

Uncertaininity in Lime

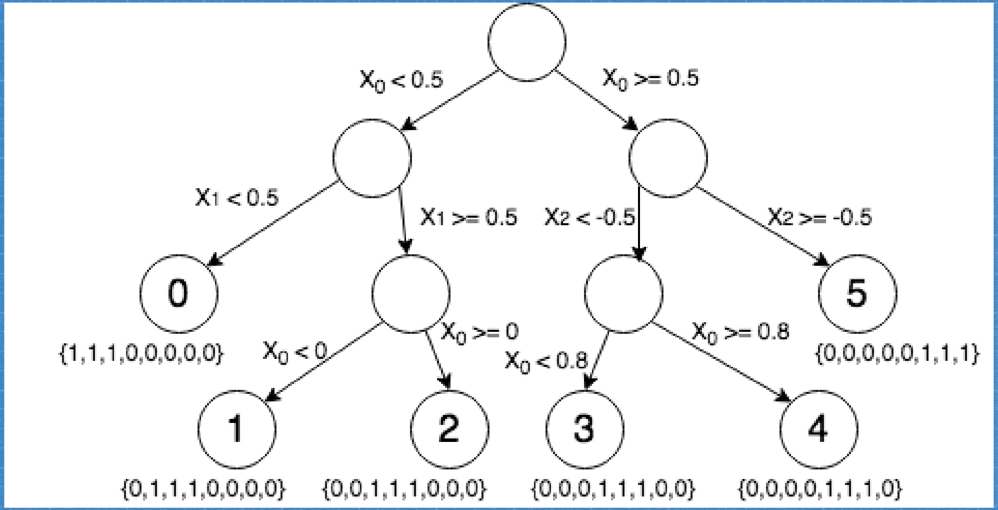
Sources of uncertainty in LIME

- Sampling variance in explaining a single data point.
- Sensitivity to choice of parameters, such as sample size and sampling proximity.
- Variation in explanation on model credibility across different data points.

Example 1: Simulation Setting

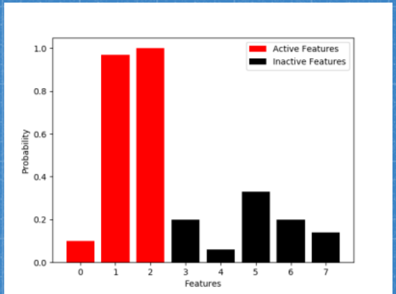
- **Data**: Eight-feature synthetic data.
  - Given the number of features N, we generate training and test data from local sparse linear models on uniformly distributed input in $[0,1]^N$.
  - To illustrate LIME's local behavior at different data points, we partition them with a known decision tree.
- **Model**: Random forest Model
- **Goal**: To illustrate the first and second source of uncertainty:
  - Randomness in the sampling procedure
  - Variation with sampling proximity.

Simulation setting: Synthetic data generated by trees

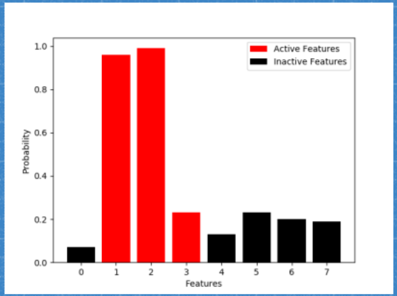


Source : Zhang et al, 2019

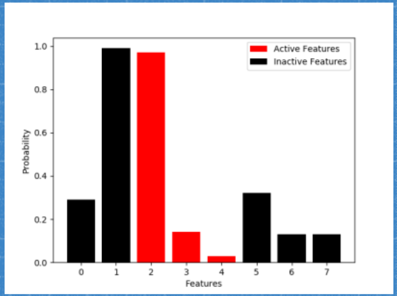
- splitting the data into six leaves for N = 8 with known coefficients, where three out of eight features have coefficients 1 in each leaf.
- Assign labels on each data point x based on a linear classifier with known coefficients
- $$y(x) = \begin{cases} 1 & x^T\beta \geq 0 \\ 0 & x^T\beta < 0 \end{cases}$$



(a) Leaf 0

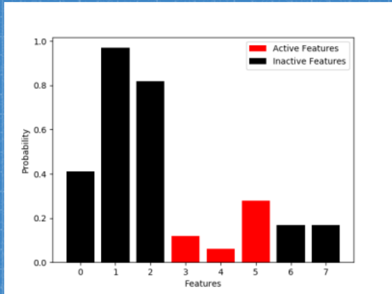


(b) Leaf 1

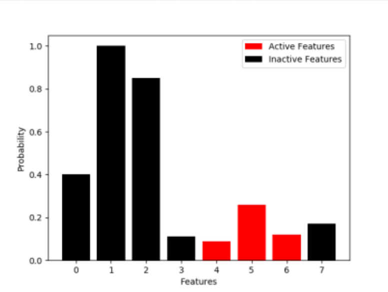


(c) Leaf 2

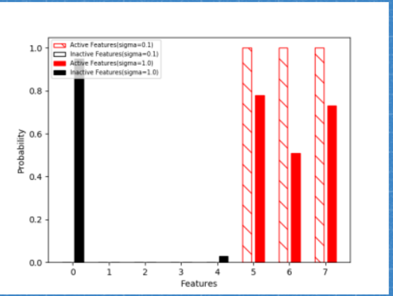
- A data point is taken from each leaf
- LIME is run 100 times on each point.
- Three feature words selected by K-LASSO
- Active features with true coefficients 1 are marked red



(d) Leaf 3



(e) Leaf 4



(f) Leaf 5

features chosen by LIME are not necessarily locally important features on each leaf. Signal from the true features is dominated by signal from the first three features used for tree splitting.

reducing the sampling proximity by a factor of ten (striped bars) which allows us to recover significant signal from the true local features and rule out the signal of feature 0 used for splitting.

Source : Zhang et al, 2019

Observations from example 1

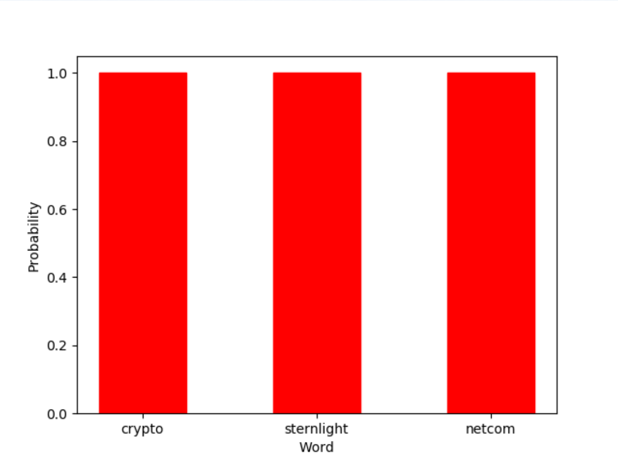
- LIME captures the signal of the first three features, which are used globally in the tree splitting of the data. Locally, however, different features are important for each individual leaf, which LIME fails to reflect.
- LIME tends to capture locally important features better with a smaller sampling proximity and pick up global features with a larger sampling proximity.

Example 2: Text Classification

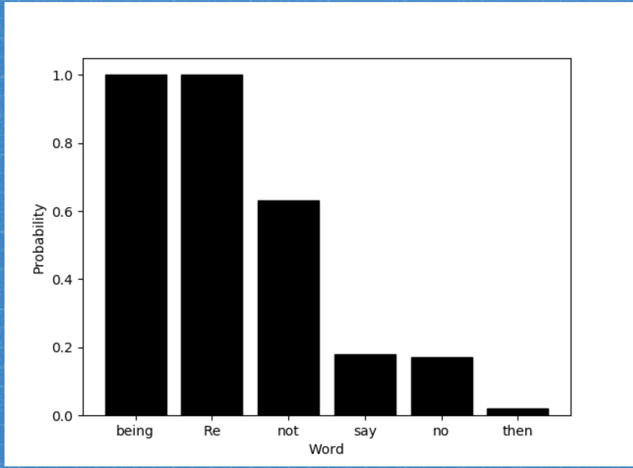
- **Data**: The 20 Newsgroup dataset is a collection of ca. 20,000 news documents across 20 newsgroups.
- **Model**: Multinomial Naive Bayes classifier

- **Goal**: To investigate variation in explanation on model credibility across different data points.
- Examples of document Classification:
  - “electronics vs. crypt”
  - “Atheism vs. Christianity”

Text Data Example 1: “electronics vs. crypt”



(a) Test Document 1



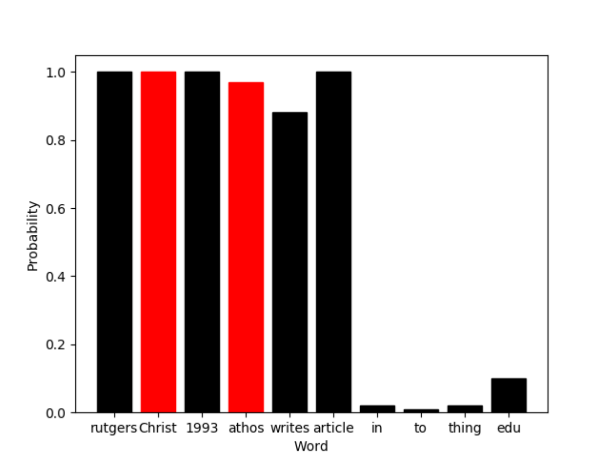
(b) Test Document 2

- LIME is run 100 times on the test document.
- three feature words selected by K-LASSO
- Informative words are marked red.

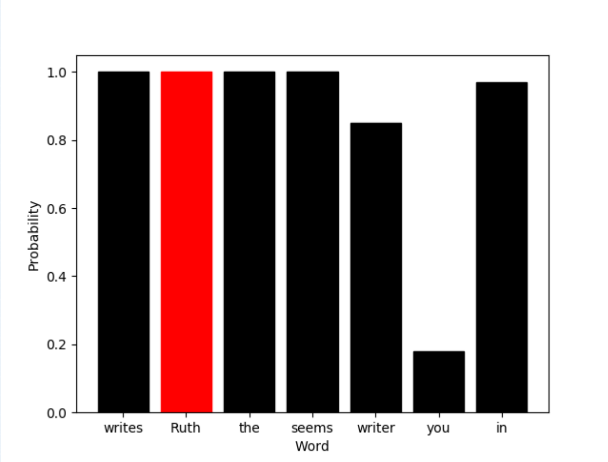
The selected feature words for the first document are consistent and meaningful, while those for the second document are not informative.

Source : Zhang et al, 2019

Text Data Example 2: “Atheism vs. Christianity”



(a) Test Document 1



(b) Test Document 2

- LIME is run 100 times on the test document.
- six feature words selected by K-LASSO
- Informative words are marked red.

Many of the frequently selected feature words are not informative.

Source : Zhang et al, 2019

Observations from example 2

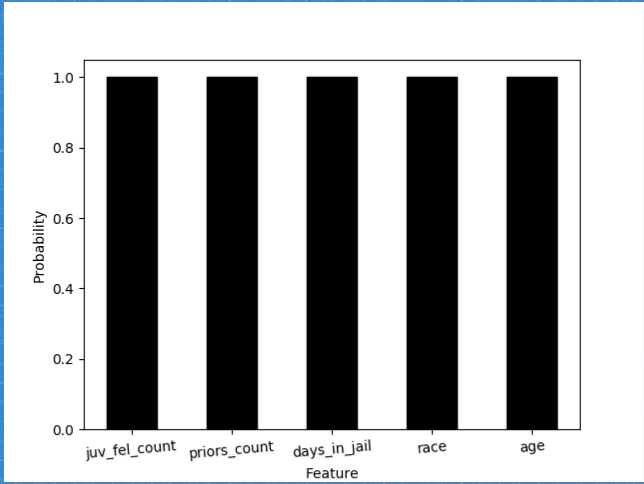
- LIME’s local explanations are not always plausible for different test documents.

- Model’s credibility, as explained by LIME, varies across different input data.

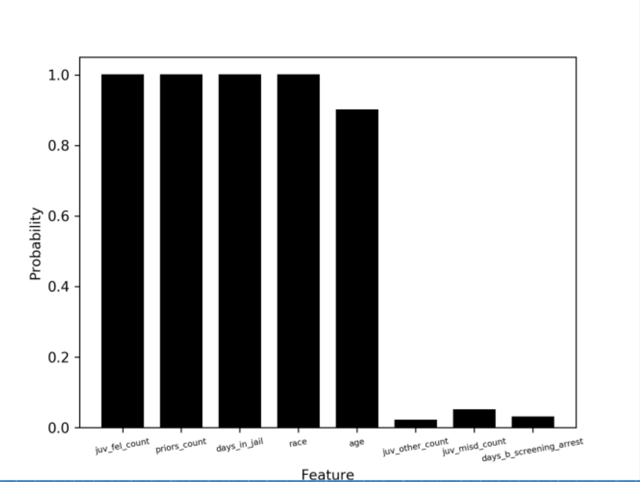
Example 3: COMPAS Example

- **Data:** subset of the COMPAS dataset collected and processed by ProPublica (Larson et al., 2016)
  - The “Correctional Offender Management Profiling for Alternative Sanctions” (COMPAS) is a risk-scoring algorithm developed by Northpointe to assess a criminal defendant’s likelihood to recidivate.
- **Model**: Random Forest classifier (“mimic model”)
- **Goal**: To show a case where LIME explanations are considered trustworthy.

COMPAS Example



(a) Sample data 1



(b) Sample data 2

- LIME is applied to two data points that are classified as “high risk” by COMPAS.

- LIME is run 50 times on the test points.
- five top features selected by K-LASSO

The features “juvenile felony count”, “priors count”, “days in jail”, “race”, and “age” are consistently selected in different trials on a single data point, as well as for two different data points.

Source : Zhang et al, 2019

Observation's from example 3

- consistent explanation results on different test data points.
  - there is little variation in the selection of important features in different trials on the same data point
  - explanation is consistent for different data points, since the same features are selected for the two different data points, including race and age.

- Further analysis using LIME suggests that the mimic model is using demographic properties

Summary

Explanation methods for black-box models may themselves contain uncertainty that calls into question the reliability of the black-box predictions and the models themselves.

References

- https://christophm.github.io/interpretable-ml-book/lime.html
- https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

Thanks!

Name: Rahul Poovassery
Matriculation No.: 229295
Supervisor: Chiara Balestra