Evaluating XAI Techniques on Interpretability Across Image, Text, and Tabular Data

By Taylor Lucero

XAI Principles

Trust

The confidence that users, including stakeholders and end-users, have in the network's predictions, decisions, and explanations.

This is often related to the Fidelity of a model or technique, which refers to the accuracy or performance of the applied method

High stake domains require a higher level of Trust and Fidelity. For example the medical domain.

Transparency

“Transparency is, roughly, a property of an application. It is about how much it is possible to understand about a system's inner workings “in theory”. It can also mean the way of providing explanations of algorithmic models and decisions that are comprehensible for the user.”

Neural Networks are often considered opaque as most of their inner workings are not easily understandable.

Interpretability VS Explainability

Interpretability	Explainability
Involves taking a stakeholder based approach on making information comprehensible	Involves providing clear reasoning on a model's prediction.
"Do I understand the presented information, and then how can I apply it?"	" Why did the model come to this decision?"
Connected to Transparency and Simplicity of a Model.	Generally requires tools/techniques to translate the prediction in relatable term.
Main Focus: To make a model more understandable for users.	Main Focus: Show how a prediction was reached regardless of transparency.

XAI Techniques

Model:
Convolutional
Neural Network
(CNN)

Model:
Long-short Term
Model
(LSTM)

Model:
Standard Artificial
Neural Network

Image

Rule Extraction
CounterFactual
Grad-Cam
Activation-Maximization

Text

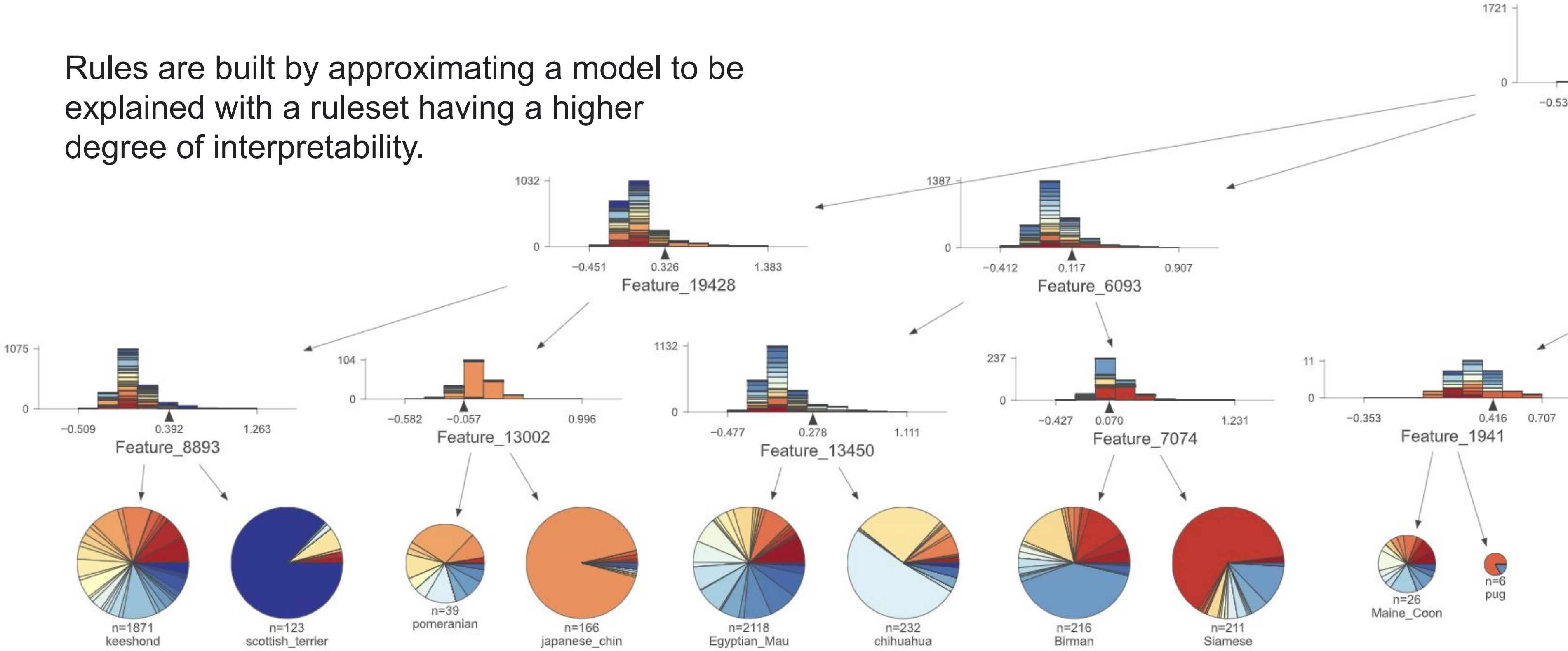
Rule Extraction
CounterFactual
Lime

Tabular

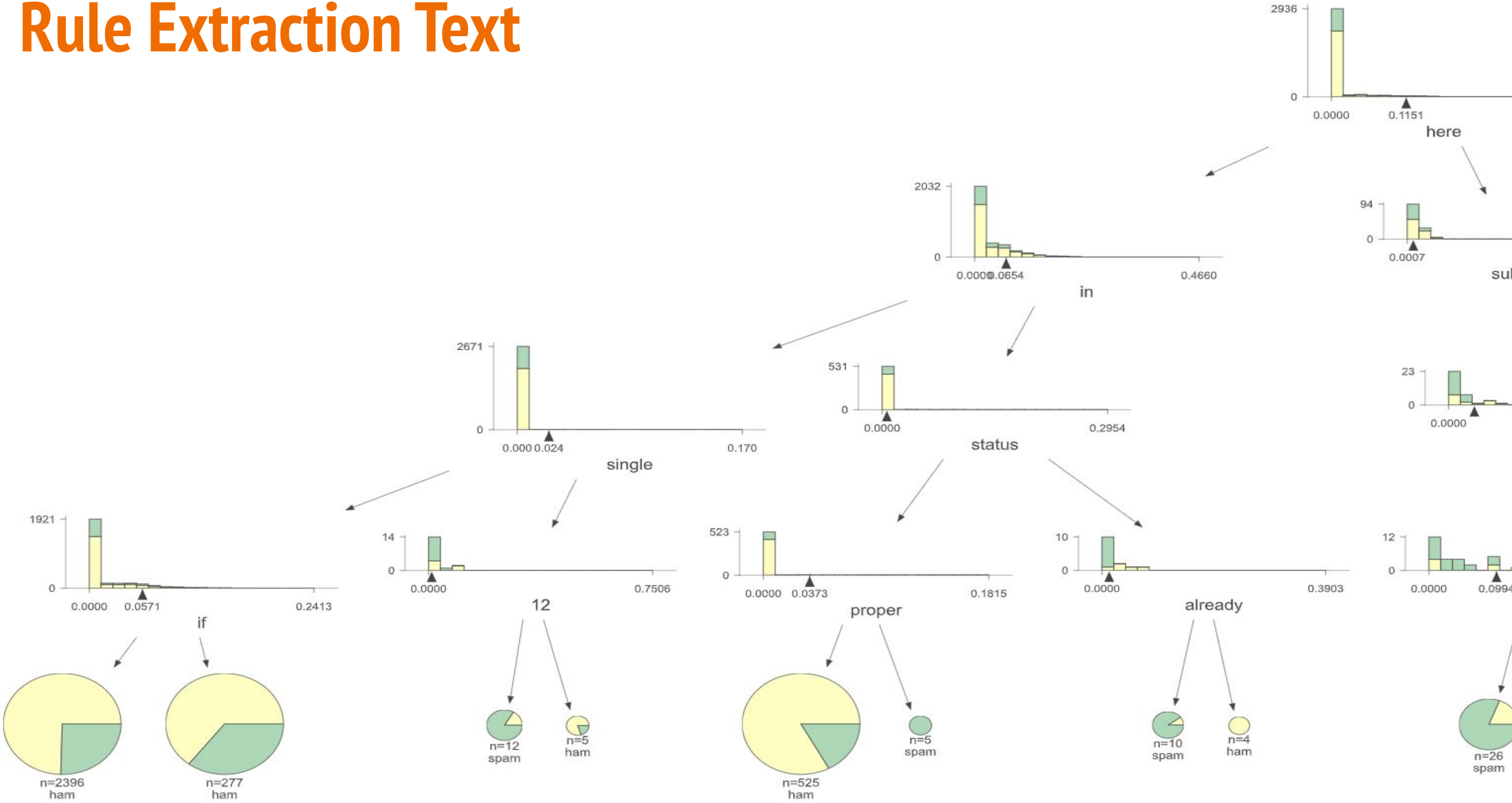
Rule Extraction
CounterFactual
PDPs
VECs

Rule Extraction Images

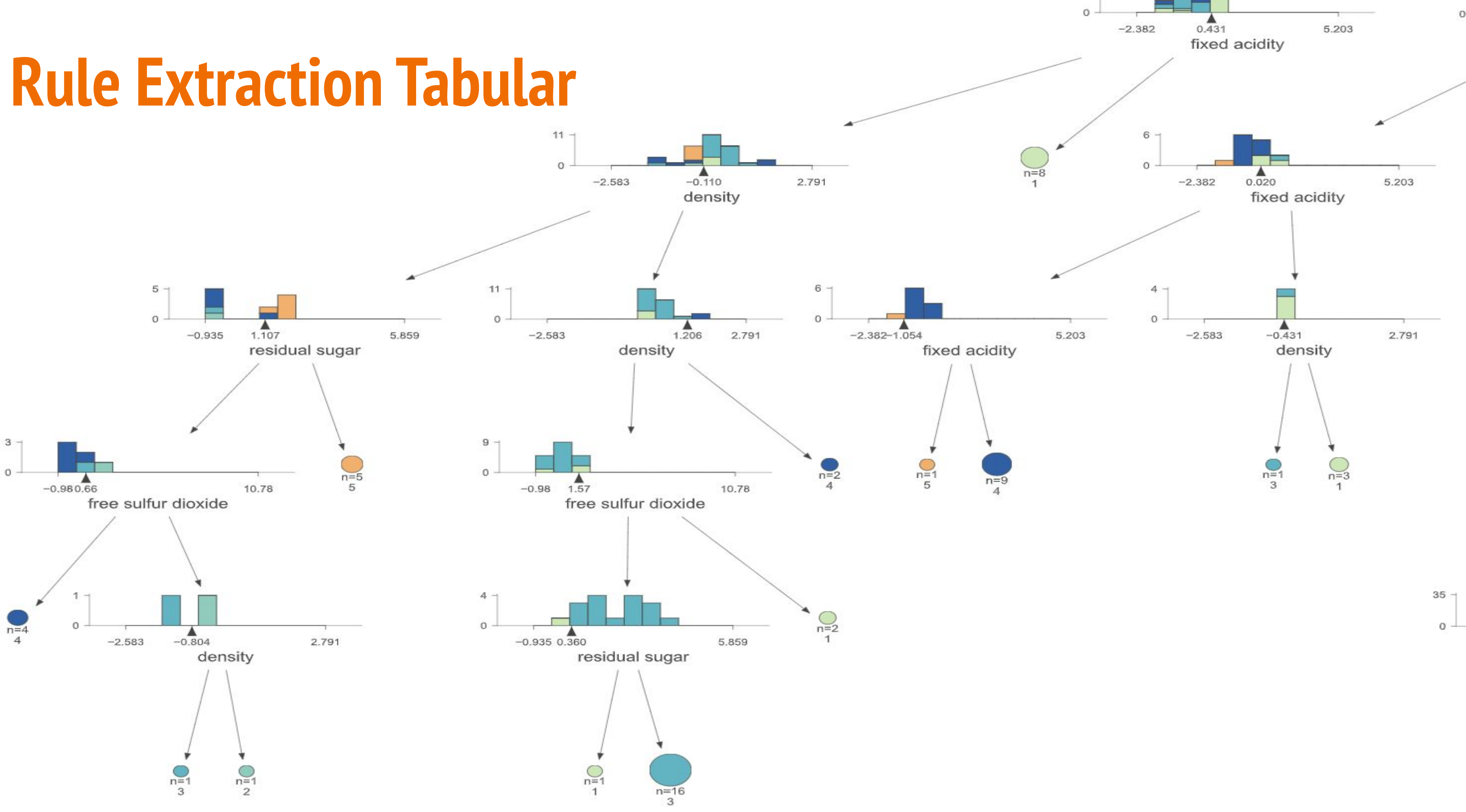
Rules are built by approximating a model to be explained with a ruleset having a higher degree of interpretability.



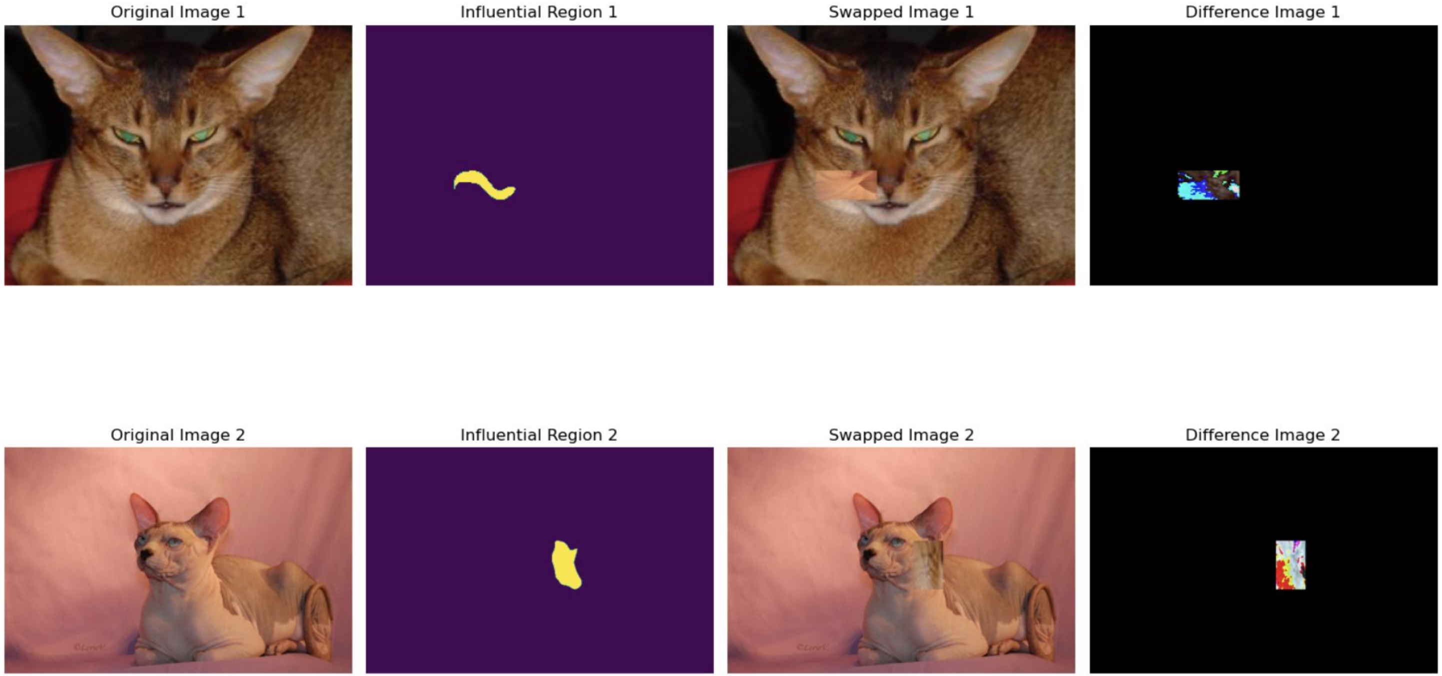
Rule Extraction Text



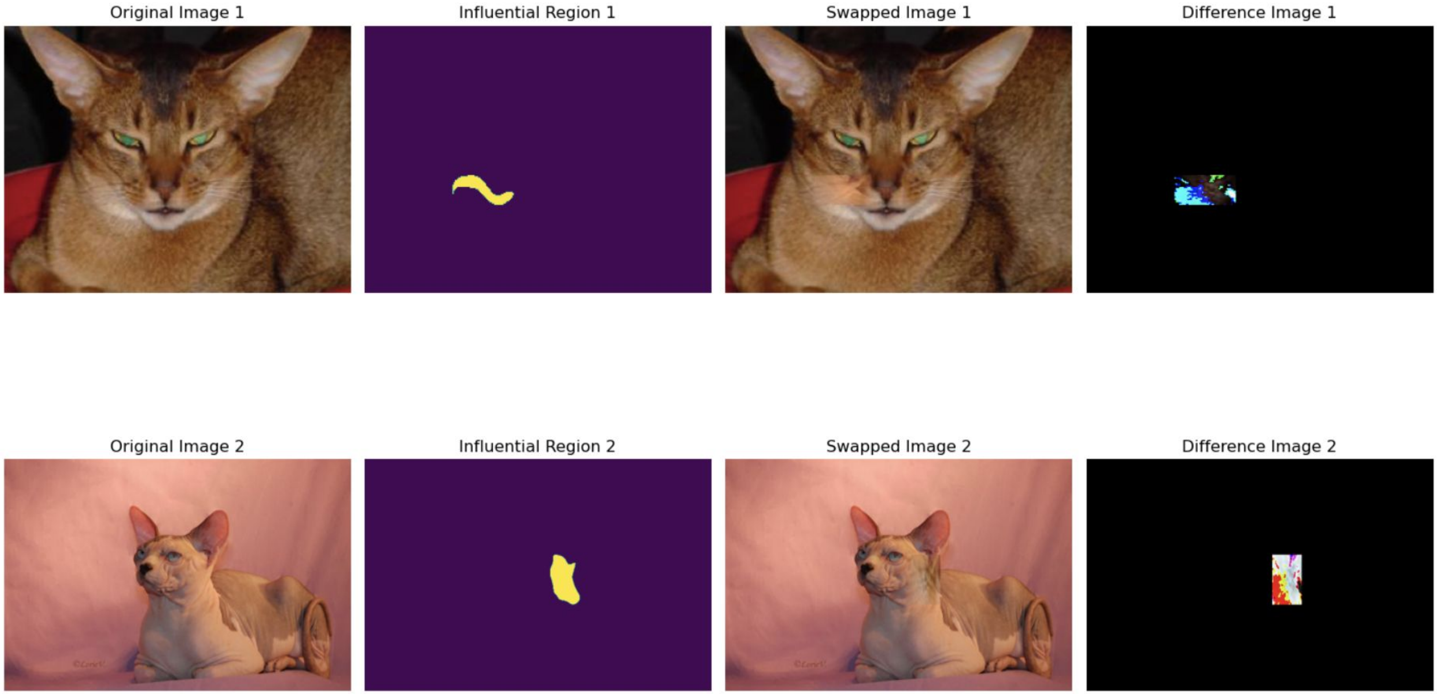
Rule Extraction Tabular



Counterfactuals: Images



Counterfactuals: Images



Counterfactuals: Text

Original: Subject: enron methanol ; meter # : 988291 this is a follow up to the note i gave you on monday , 4 / 3 / 00 { preliminary flow data provided by daren } .please override pop ' s daily volume { presently zero } to reflect daily activity you can obtain from gas control . this change is needed asap for economics purposes .

Counterfactual 2: subject : enron methanol ; product # : 988291 this is a follow up to the note carter gave you on monday, 4 / 3 / 00 { all contact documents provided by source }. please override pop's full volume { if applicable } to reflect daily activity you can obtain from gas control. this change not needed asap for economics purposes.

Counterfactuals : Tabular

Original Data

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	color	quality
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	0	0

Table 10.1: Original Data

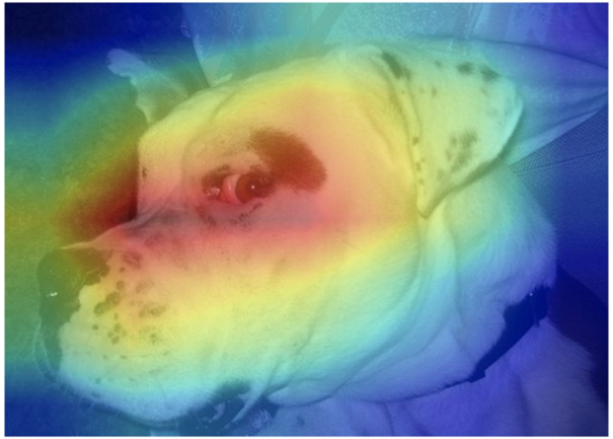
Diverse Counterfactual set (new outcome: 5)

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	color	quality
-	-	-	-	-	-	7.0	0.989	3.85	-	-	1.0	5.0
-	-	-	20.8	-	-	28.1	-	3.24	-	-	1.0	5.0
-	-	-	22.9	-	-	-	-	-	1.98	13.9	-	5.0
-	1.22	-	25.5	-	-	-	-	3.51	-	13.7	-	5.0

Table 10.2: Diverse Counterfactual Set

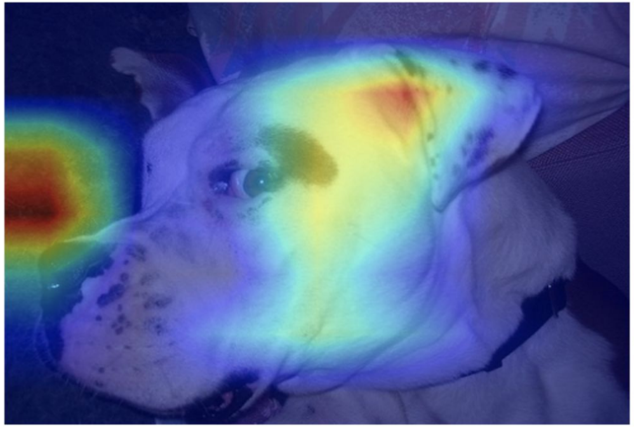
Grad-Cam

Activation Map for Layer: 0.7.2.conv2

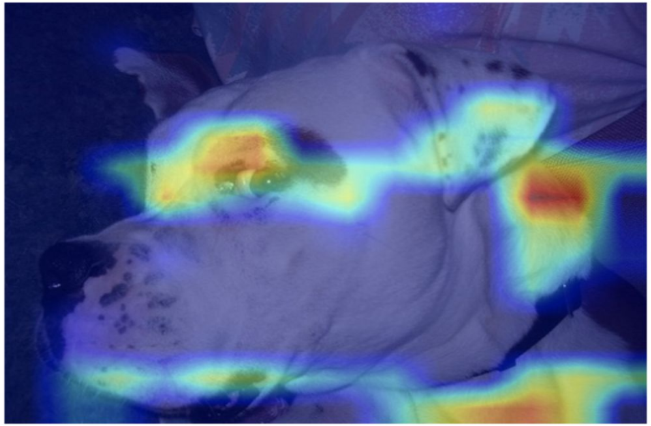


Activation Map for Layer: 0.7.2.conv1

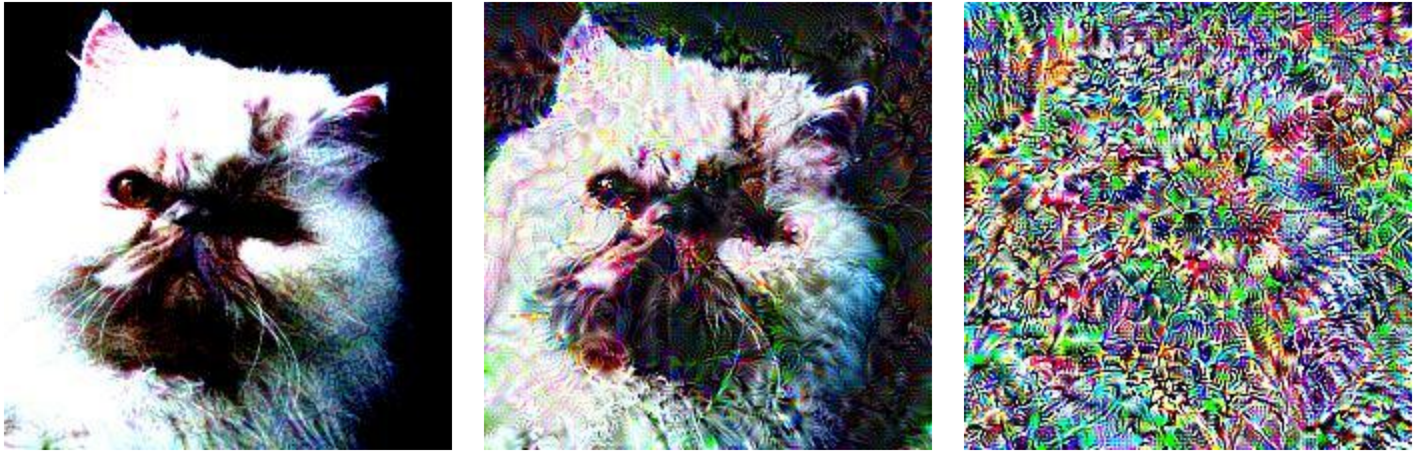
Activation Map for Layer: 0.7.1.conv2



Activation Map for Layer: 0.6.3.conv2

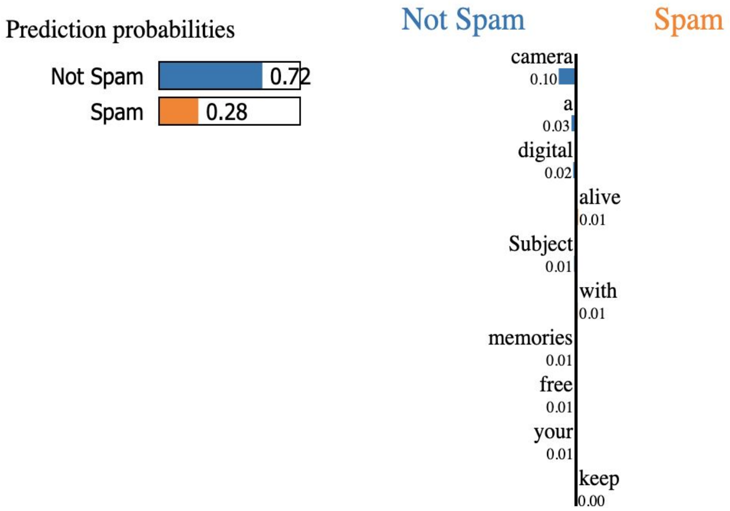


Activation Maximization



Lime for Attribution Score (Text)

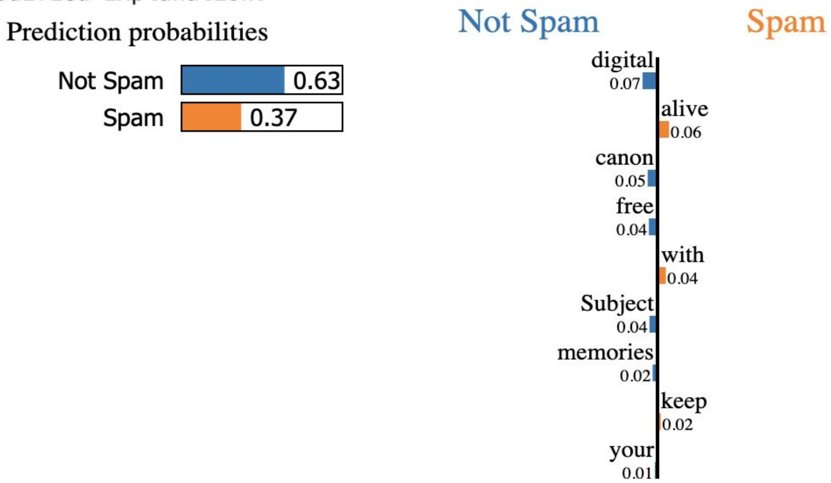
```
Subject: keep your memories alive with a free canon digital camera !
>
```



Text with highlighted words

Subject: keep your memories alive with a free canon digital camera !

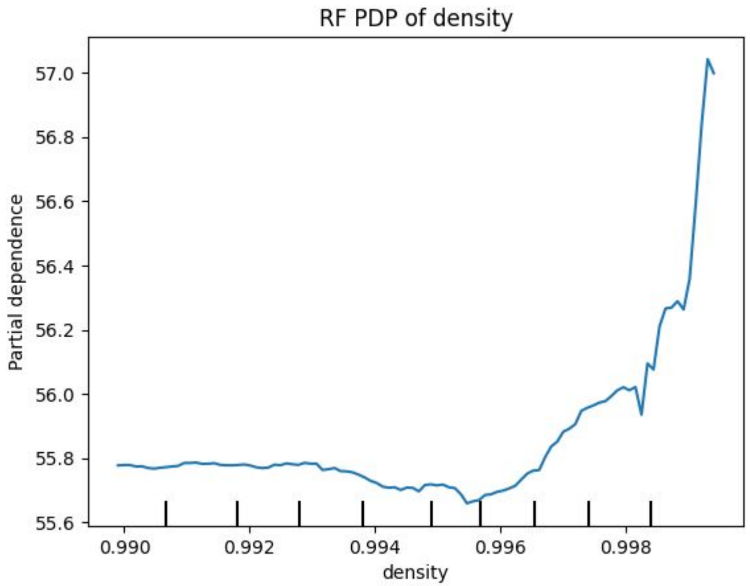
```
Modified Instance:
Subject: keep your memories alive with free canon digital !
>
Modified Explanation:
```



Text with highlighted words

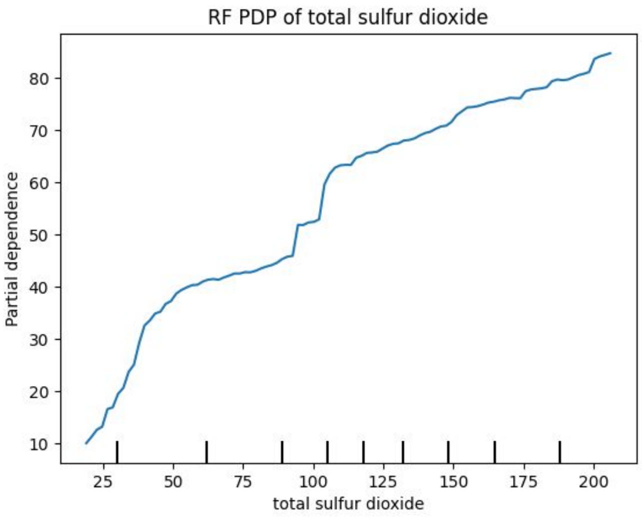
Subject: keep your memories alive with free canon digital !

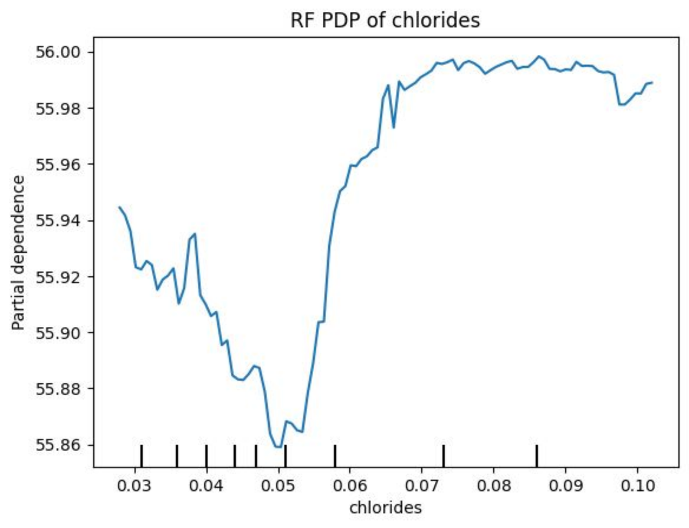
Partial Dependence Plots



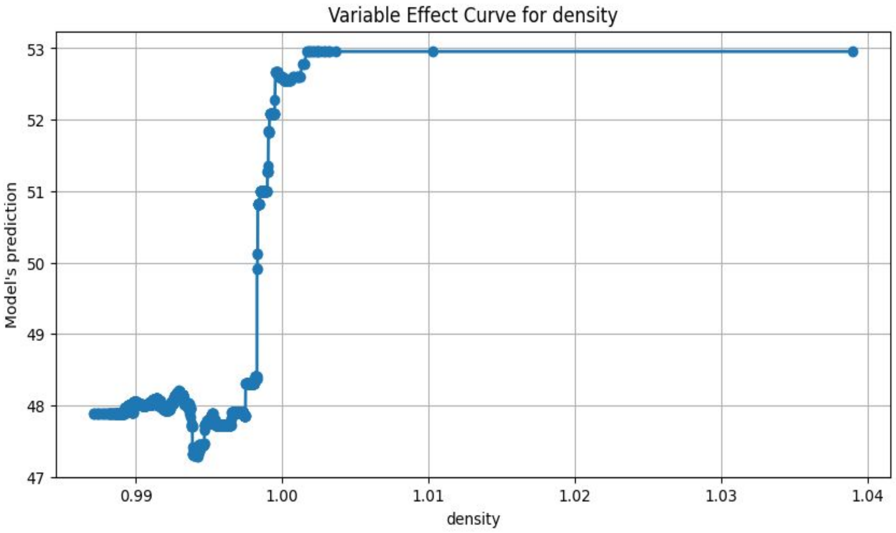
- Shows more general trends without many fluctuations.
- Visualizes the marginal effect of one or two features on the predicted outcome of a model

Partial Dependence Plots



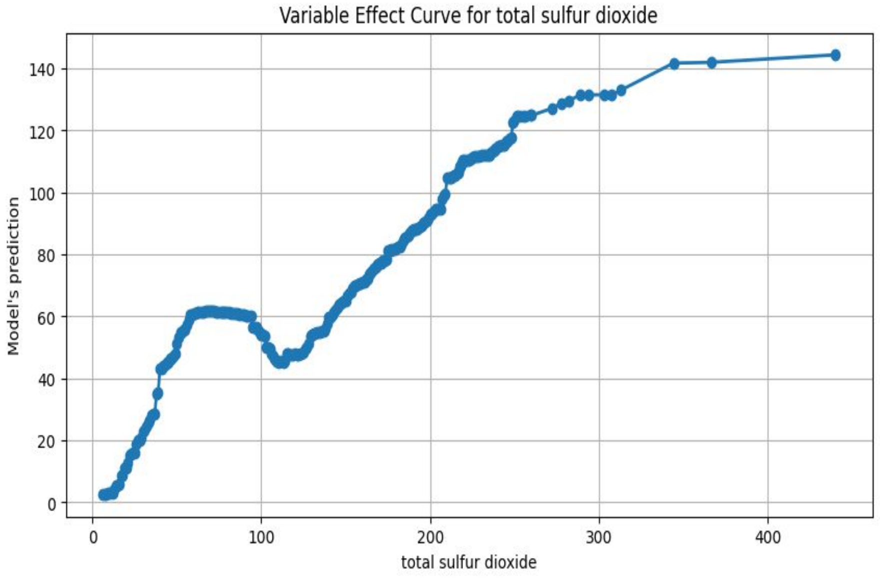


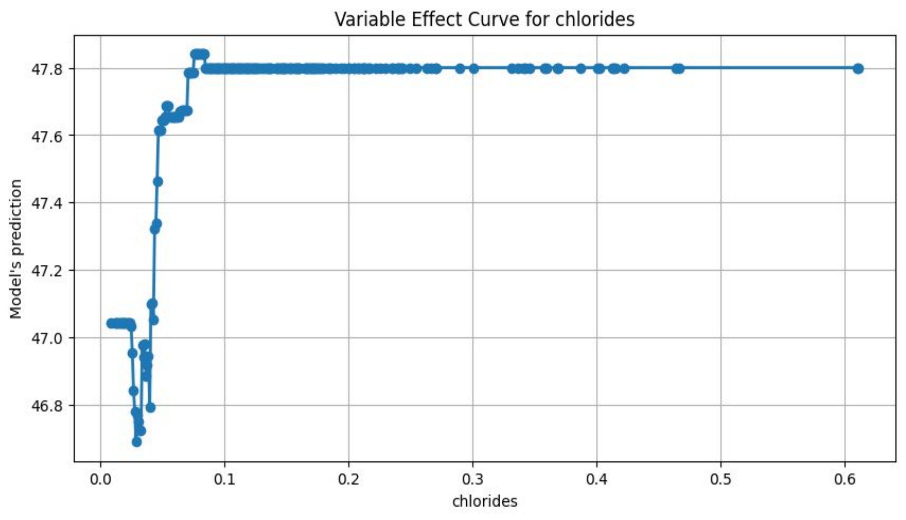
Variable Effect Curves



- VECs can capture more intricate local patterns
- Help shows the relationship between specific features (or variables) and the model predictions

Variable Effect Curves





Overall

- The rule extractions were difficult to comprehend visually, this improved only with the help of treeviz allowing for a more interactive notebook and providing the deciding feature for each node. Specifically, rule extraction with images is difficult based on how pixels are treated in the CNN.
- Partial Dependence Plots and Variable Effect Curves provide a more impactful information when visualized together than by themselves.
- Counterfactuals for image data based on the technique used is difficult to use as it can take portion of images and swap them in incorrect or unnatural positions.
- Counterfactuals for text and tabular, was more helpful in understanding important features that have an impact on the models predictions. However, for both, unnatural instances can be produced.