Is Your ML Model Trustworthy?

María Grandury

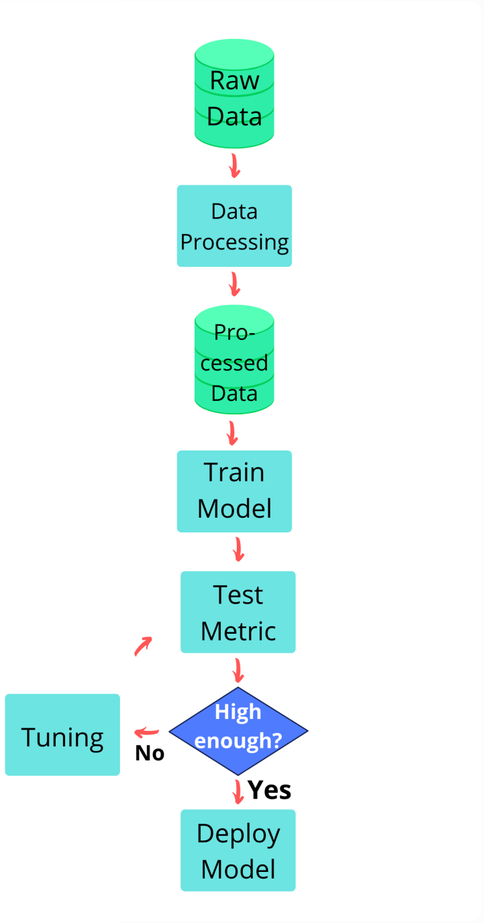
mariagrandury.github.io

@mariagrandury

About me

- 💡 Machine Learning Research Engineer
- 🎯 #NLP, AI Robustness & Explainability (#XAI)
- 🎓 Mathematician & Physicist
- 🤖 Trusted AI @neurocat.ai
- 🚀 Founder @NLP_en_ES 🤗
- ⚡ Core Team @WAIRobotics

Functionality



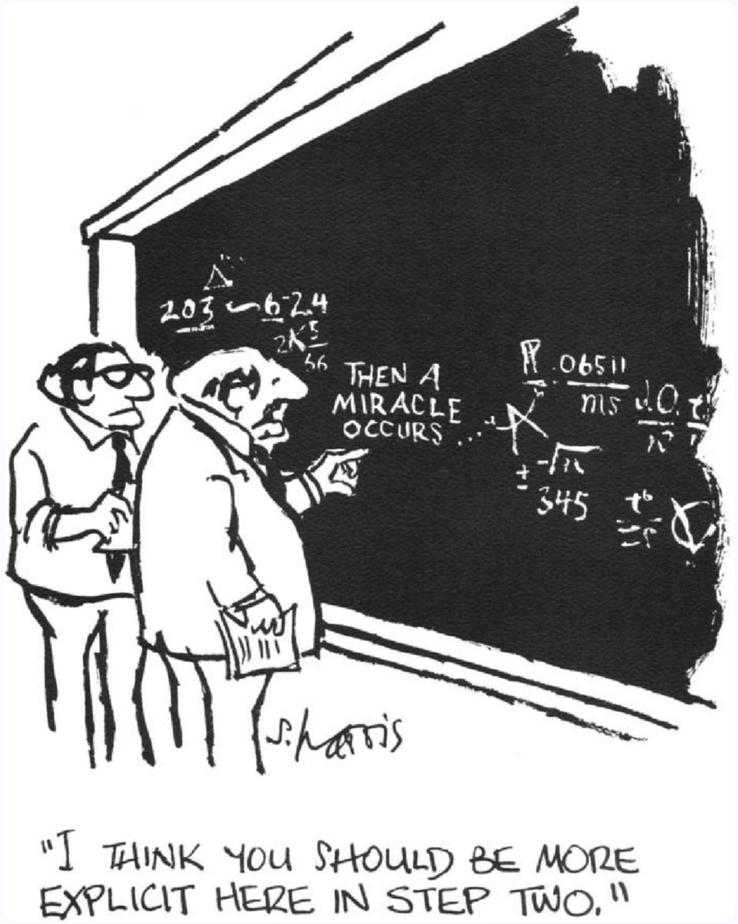
Performance Metrics

- Classification: Accuracy, F1-score, AUC
- Regression: MSE, MAE, R^2
- Ranking: Mean Reciprocal Rank
- NLP: BLEU, ROUGE, Perplexity
- GANs: Inception score, Frechet Inception distance

THE Question

- "Is my performance metric high enough?"
- ☑ "Is my model trustworthy enough?"

Explainability



Explainability

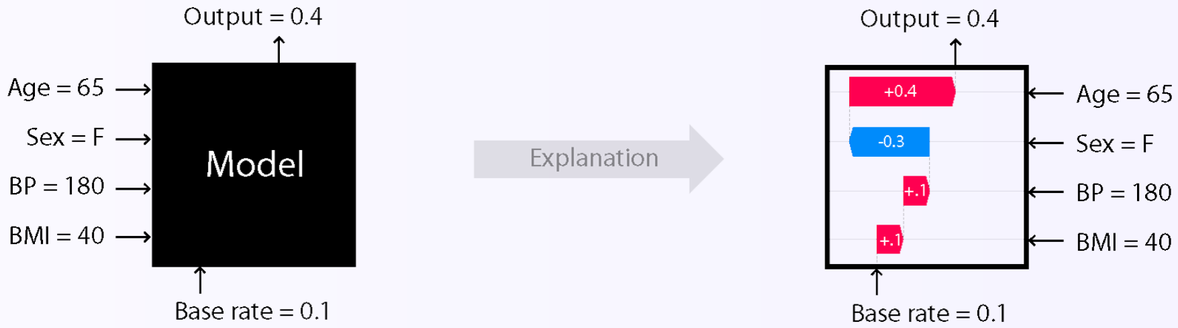
XAI plays a crucial role in sensitive domains.

Explaining ML Models with SHAP

📄 Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions", 2017, NIPS

☆ github.com/slundberg/shap (12.9k)

- SHapley Additive exPlanations
- Game theoretic approach
- Explains the output prediction by computing the contribution of each feature to it



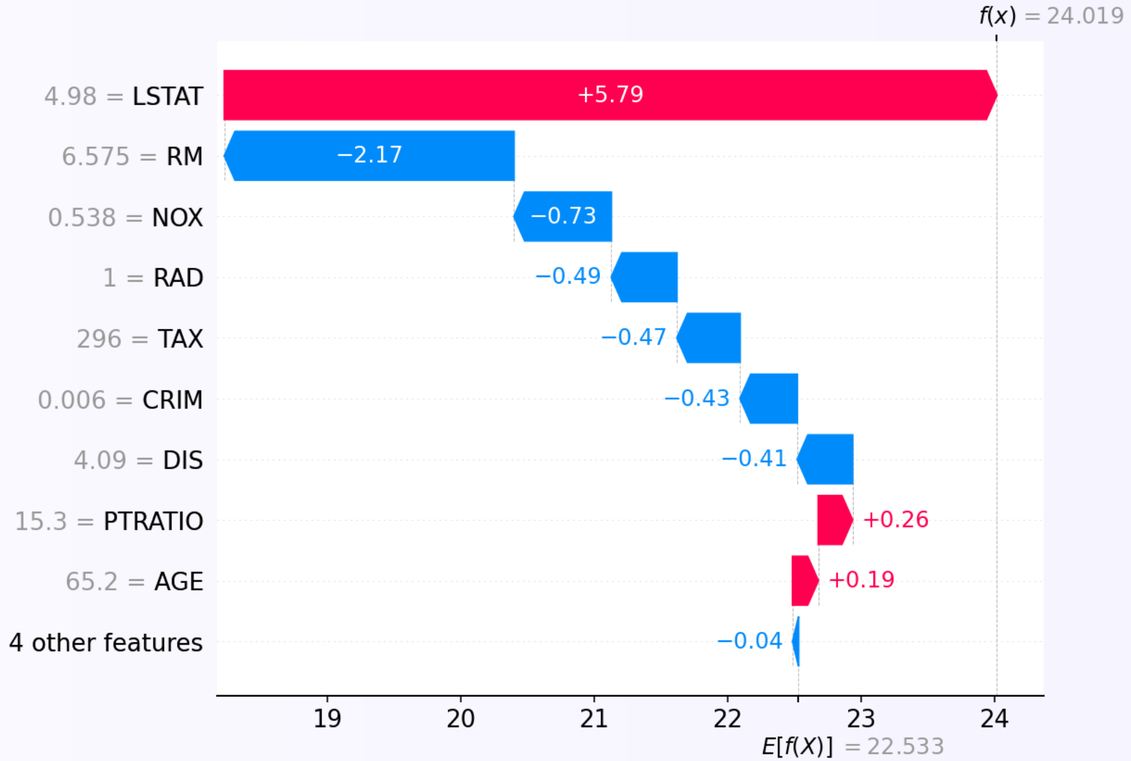
Explainability with SHAP - Tabular Data

The features contribute to push the model output from the base value to the model output:

- Push the prediction higher
- Push the prediction lower

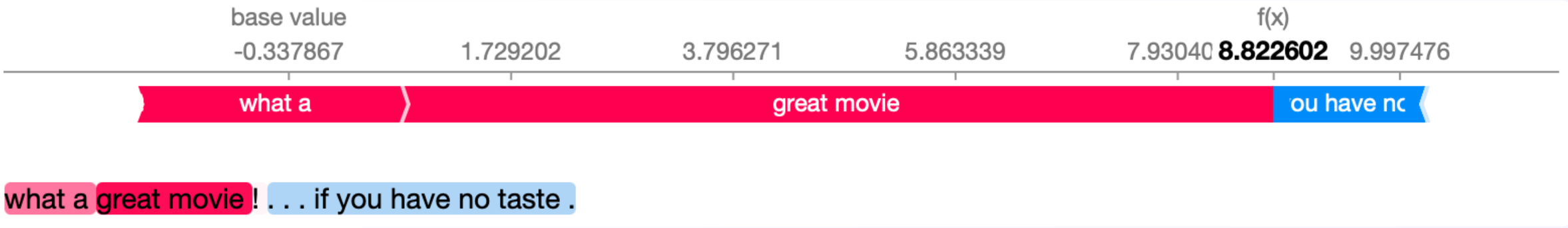
Example:

- Boston Housing data set
  - LSTAT: % lower status of the population
  - RM: average number of rooms per residence
- Regression Model (XGBoost)



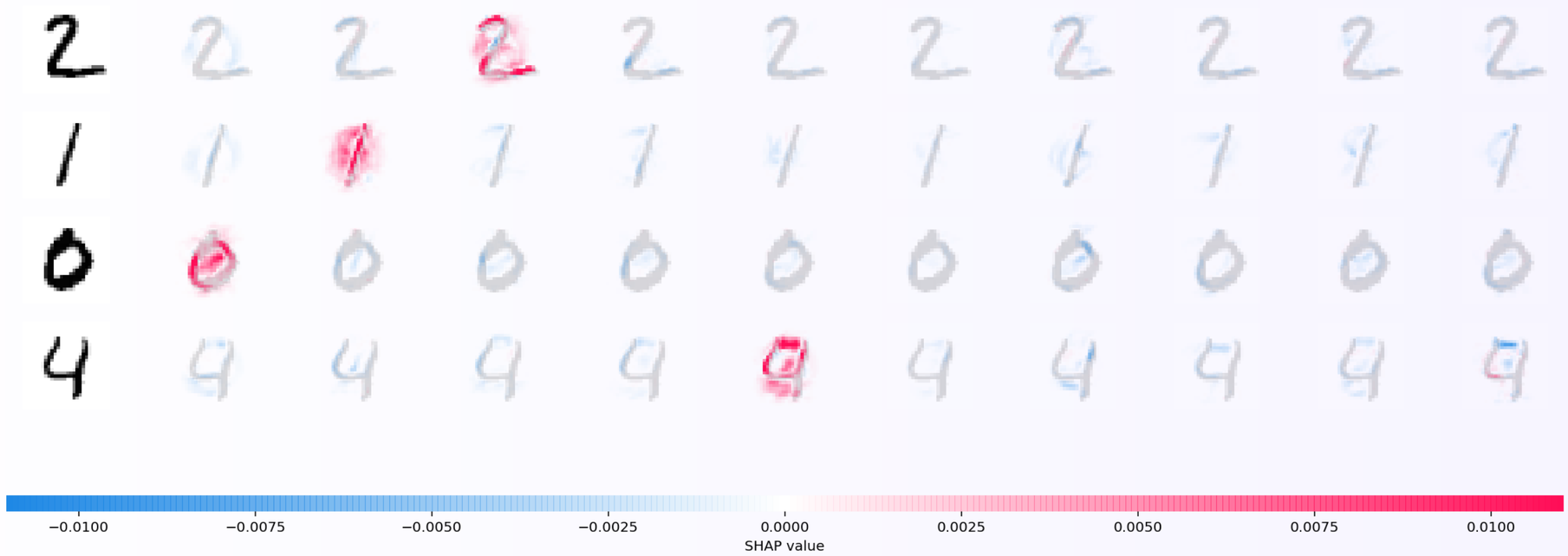
Explainability with SHAP - NLP

- Force plot
- Lundberg, Lee et al., "Explainable ML predictions for the prevention of hypoxaemia during surgery", 2018, doi: 10.1038/s41551-018-0304-0
- IMDb movie review data set
- Sentiment Analysis
- 🤗 Transformers
- Explanation for the POSITIVE output class

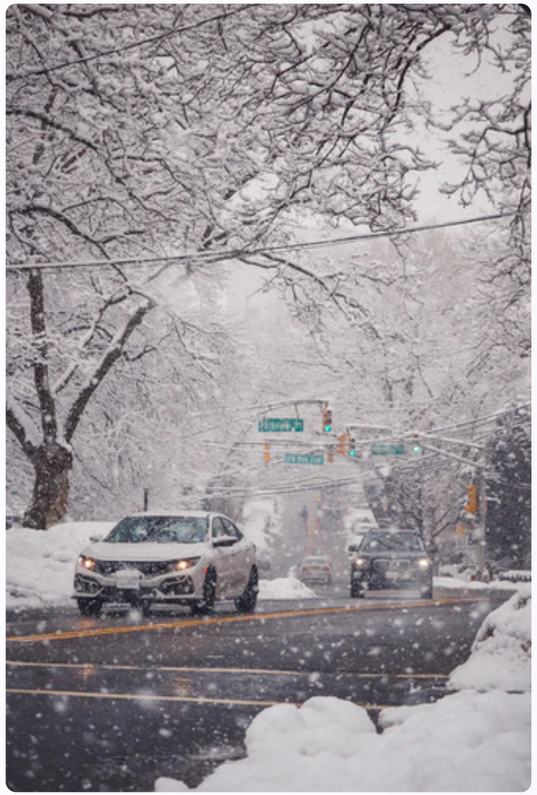


Explainability with SHAP - Computer Vision

- MNIST data set
- Classification Model



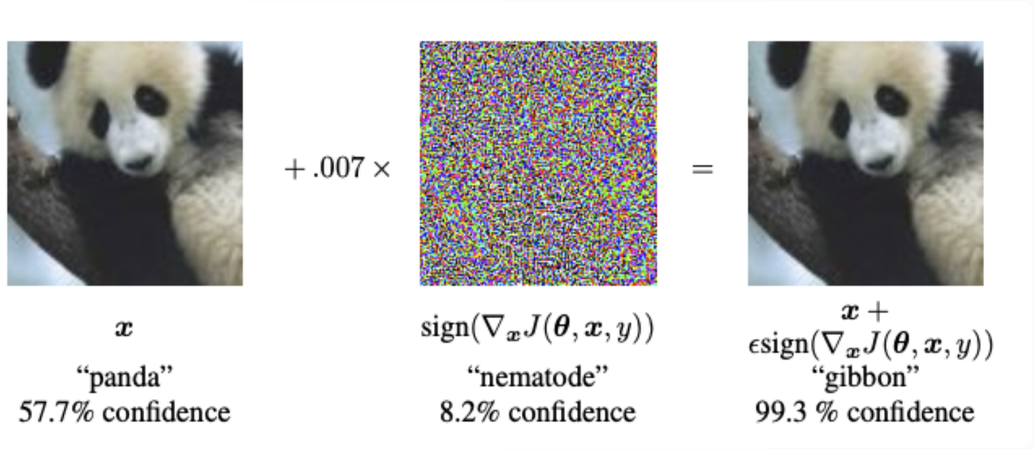
Robustness

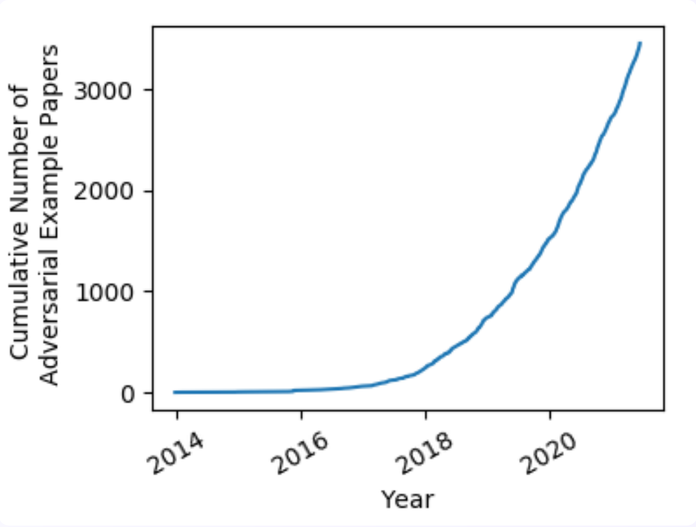


Adversarial Attacks

📄 Szegedy, “Intriguing properties of neural networks”, 2013, arXiv:1312.6199 [cs.CV]

📄 Goodfellow, Shlens and Szegedy, "Explaining and Harnessing Adversarial Examples", 2014, arXiv:1412.6572 [stat.ML]





A Complete List of All (arXiv) Adversarial Example Papers by Nicholas Carlini.

Adversarial Attacks with CleverHans

Papernot, Faghri, Carlini, Goodfellow et al., "Technical Report on the CleverHans v2.1.0 Adversarial Examples Library", 2018, arXiv:1610.00768 [cs.LG]

☆ github.com/cleverhans-lab/cleverhans (5.1k)

- Attacks: FGSM Attack, Carlini Wagner Attack
- Defenses: Resampling

CleverHans Blog

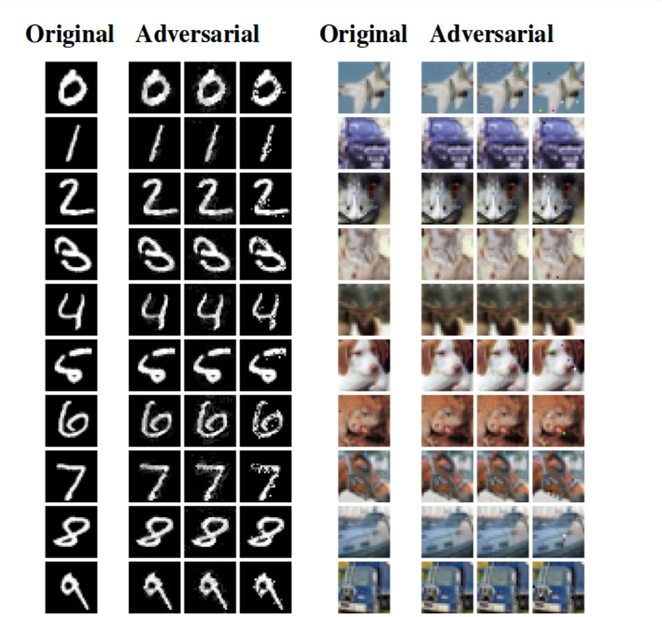


Fig. 1. An illustration of our attacks on a defensively distilled network. The leftmost column contains the starting image. The next three columns show adversarial examples generated by our L_2, L_∞, and L_0 algorithms, respectively. All images start out classified correctly with label l, and the three misclassified instances share the same misclassified label of $l+1 \pmod{10}$. Images were chosen as the first of their class from the test set.

Adversarial Attacks with TextAttack

📄 Morris et al., "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP", 2020, arXiv:2005.05909 [cs.CL]

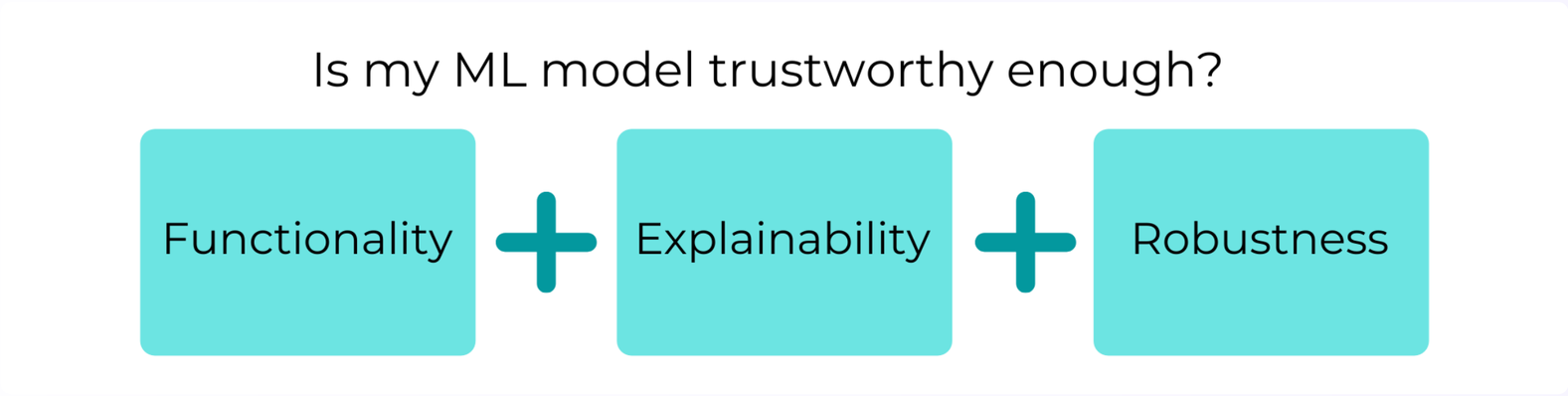
☆ github.com/QData/TextAttack (1.5k)

Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Positive (77%)**
Adversarial example [Visually similar]	**Aonnoisseurs** of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (52%)**
Adversarial example [Semantically similar]	Connoisseurs of Chinese **footage** will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (54%)**

MLOps Workflow

MLOps Tools: MLflow, Airflow, Neptune, Kubeflow, MLrun…

- Add one step before deployment!



Thank you!

Let’s shape the future of AI Quality!

aidkit.ai

Slides can be found at github.com/mariagrandury

More Resources

Explainability:

- C. Molnar, "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019, christophm.github.io/interpretable-ml-book
- A. Saucedo, "Guide towards algorithm explainability in machine learning", talk at PyData London 2019
- S. Lundberg, "Explainable Machine Learning with Shapley Values", talk at #H2OWorld 2019
- A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in ML", 2020, doi: 10.1145/3376898
- V. Dignum, "The Mith of Complete AI Fairness", 2021, arXiv:2104.12544v1 [cs.CY]

Adversarial Attacks:

- "Attacking Machine Learning with Adversarial Examples", OpenAI Blog Post
- I. Goodfellow, "Adversarial Examples and Adversarial Training", lecture at Stanford University
- J. Morris, "TextAttack: A Framework for Data Augmentation and Adversarial Training in NLP", talk at dair.ai

Code: SHAP & Tabular Data

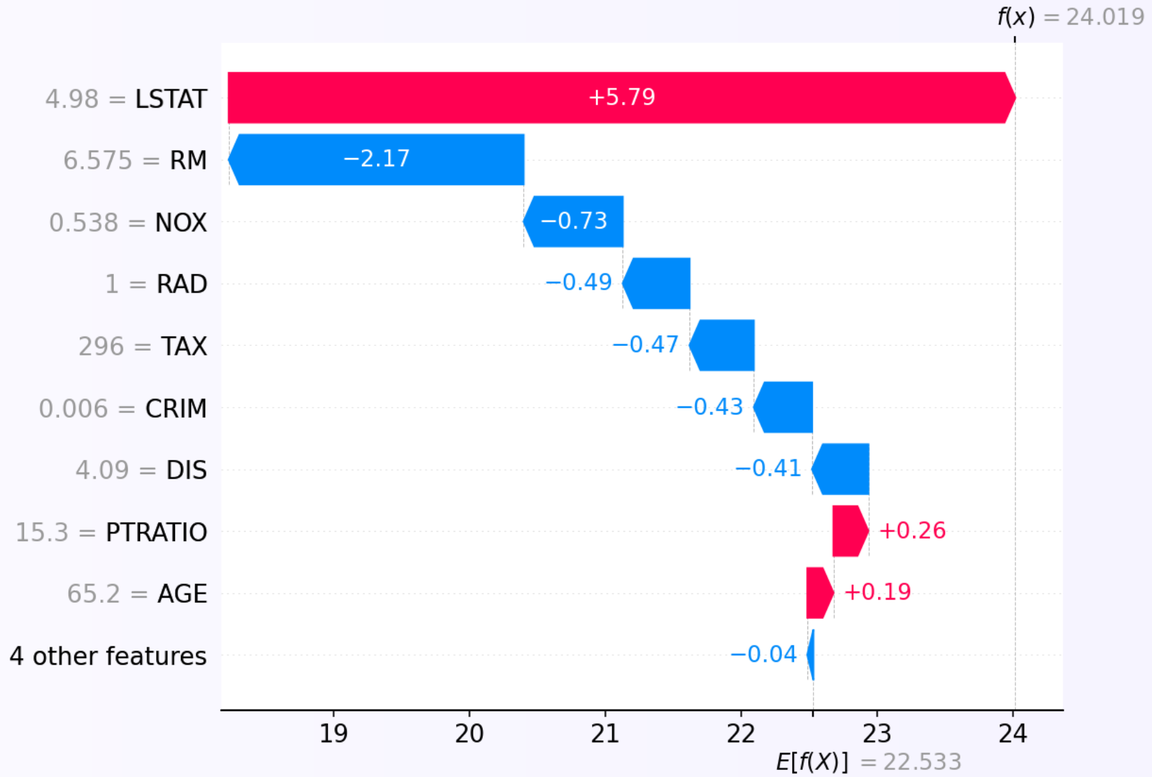
```
!pip install shap

import xgboost
import shap

# train an XGBoost model
X, y = shap.datasets.boston()
model = xgboost.XGBRegressor().fit(X, y)

# explain the model's predictions using SHAP
explainer = shap.Explainer(model)
shap_values = explainer(X)

# visualize the first prediction's explanation
shap.plots.waterfall(shap_values[0])
```



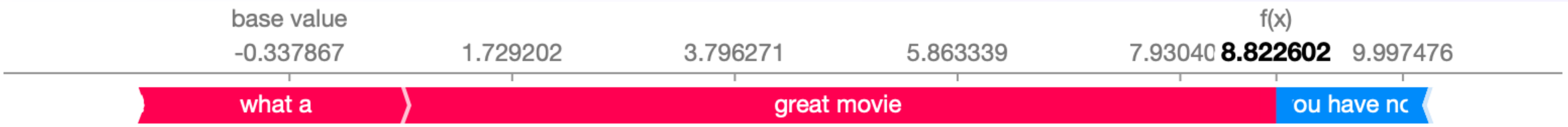
Code: SHAP & NLP

```
import transformers
import shap

# load a transformers pipeline model
model = transformers.pipeline('sentiment-analysis', return_all_scores=True)

# explain the model on two sample inputs
explainer = shap.Explainer(model)
shap_values = explainer(["What a great movie! ...if you have no taste."])

# visualize the first prediction's explanation for the POSITIVE output class
shap.plots.text(shap_values[0, :, "POSITIVE"])
```



what a great movie ! . . . if you have no taste .

Code: SHAP & CV

```
import shap
import numpy as np

# select a set of background examples to take an
# expectation over
background = x_train[np.random.choice(
  x_train.shape[0], 100, replace=False
)]

# explain predictions of the model on four images
e = shap.DeepExplainer(model, background)
shap_values = e.shap_values(x_test[1:5])

# plot the feature attributions
shap.image_plot(shap_values, -x_test[1:5])
```

Code: Cleverhans

```
# Train model with adversarial training
for epoch in range(FLAGS.nb_epochs):
  # keras like display of progress
  progress_bar_train = tf.keras.utils.Progbar(60000)
  for (x, y) in data.train:
      if FLAGS.adv_train:
          # Replace clean example with adversarial example for adversarial training
          x = projected_gradient_descent(model, x, FLAGS.eps, 0.01, 40, np.inf)
      train_step(x, y)

# Evaluate on clean and adversarial data
progress_bar_test = tf.keras.utils.Progbar(10000)
for x, y in data.test:
  y_pred = model(x)
  test_acc_clean(y, y_pred)

  x_fgm = fast_gradient_method(model, x, FLAGS.eps, np.inf)
  y_pred_fgm = model(x_fgm)
  test_acc_fgsm(y, y_pred_fgm)

  x_pgd = projected_gradient_descent(model, x, FLAGS.eps, 0.01, 40, np.inf)
  y_pred_pgd = model(x_pgd)
  test_acc_pgd(y, y_pred_pgd)
```

Code: TextAttack

```
#!/bin/bash
# how to attack a DistilBERT model fine-tuned on SST2 dataset *from the
# huggingface model hub using the DeepWordBug recipe and 10 examples

textattack attack
  --model-from-huggingface distilbert-base-uncased-finetuned-sst-2-english
  --dataset-from-huggingface glue^sst2
  --recipe deepwordbug
  --num-examples 10
```