Google Cloud

03

Interpretability of AI

Introduction to Responsible AI in Practice

In this module, you learn to ...

01 Define **interpretability** in ML

02 Discuss why **interpretability is important** and **difficult**

03 Discover some **best practices** on interpretability

04 Explore **techniques** and **tools** to study interpretability

05 **Lab: Learning Interpretability Tool** for **Text Summarization**

Topics

01	Overview of Interpretability
02	Metrics Selection
03	Taxonomy of interpretability in ML Models
04	Tools to Study Interpretability
05	Hands-on Lab

Topics

01	Overview of Interpretability
02	Metrics Selection
03	Taxonomy of interpretability in ML Models
04	Tools to Study Interpretability
05	Hands-on Lab

Interpretability relates to Google's AI Principle #4

1 Be socially beneficial

2 Avoid creating or reinforcing unfair bias

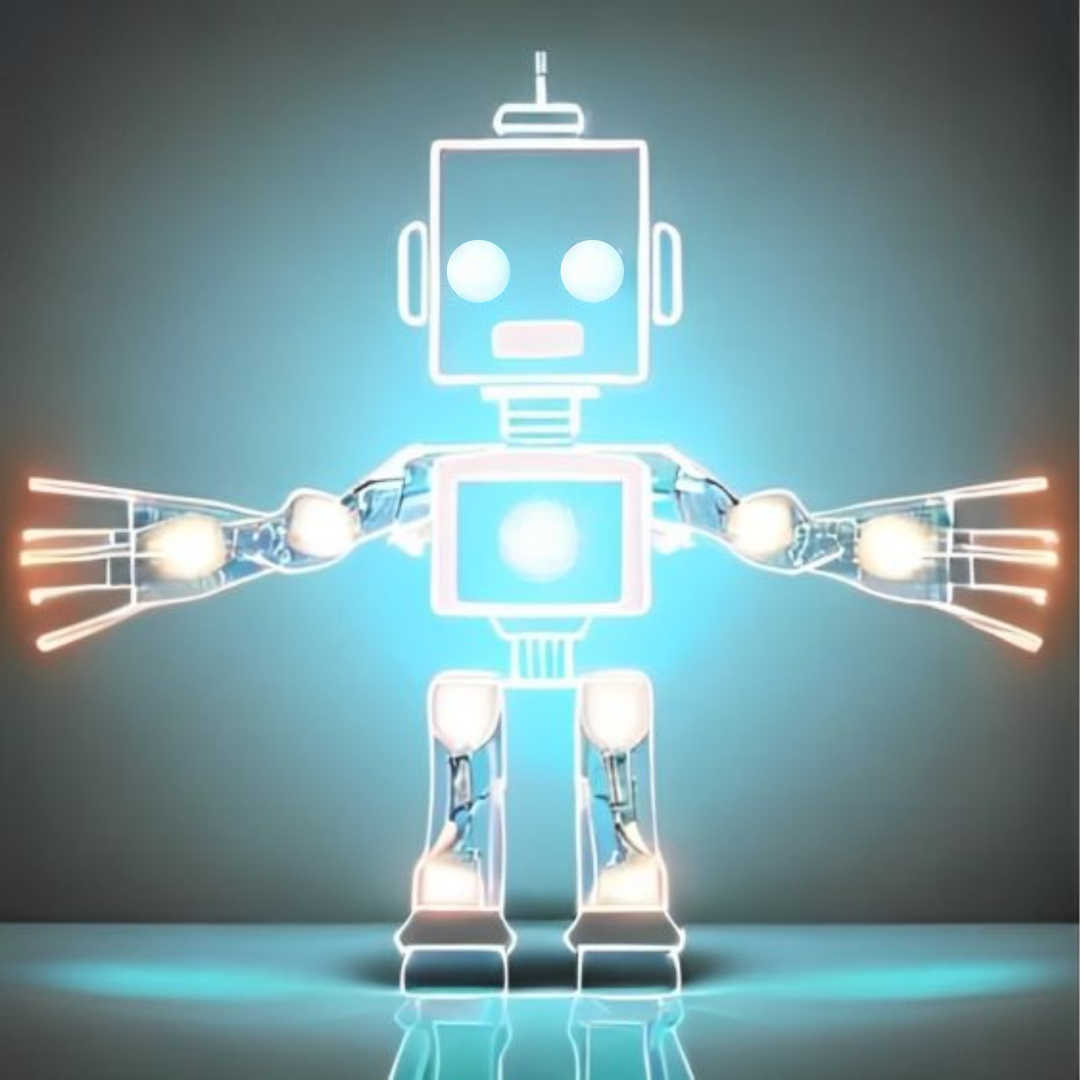
3 Be built and tested for safety

4 **Be accountable to people**

5 Incorporate privacy design principles

6 Uphold high standards of scientific excellence

7 Be made available for uses that accord with these principles



AI Interpretability

The ability to explain or to present an ML model's reasoning in understandable terms to a human.

Definition from https://developers.google.com/machine-learning/glossary

What makes a good explanation?

- Completeness
- Accuracy
- Meaningfulness
- Consistency

How does Interpretability fit with explainability?

Interpretability = Explainability ?

Why do you need Interpretability?

- Question
- Understand
- Trust

- Reflect our domain knowledge and societal values
- Provide scientists and engineers with better means
- Ensure AI systems are working as intended
- Present models to stakeholders

Why is Interpretability difficult?

Not easy for anyone

Interpretability issues apply to humans as well as AI systems—after all, it's not always easy for a person to provide a satisfactory explanation of their own decisions.

Model complexity

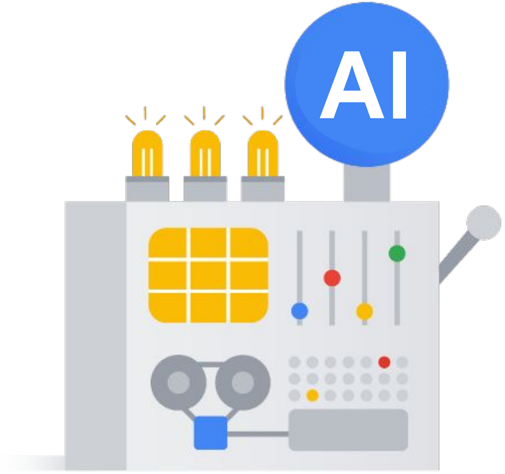
Understanding complex AI models, such as deep neural networks, can be challenging even for machine learning experts.

ML vs traditional software

Understanding and testing AI systems also offers new challenges compared to traditional software.
It is much harder to pinpoint one specific bug that leads to a faulty decision.

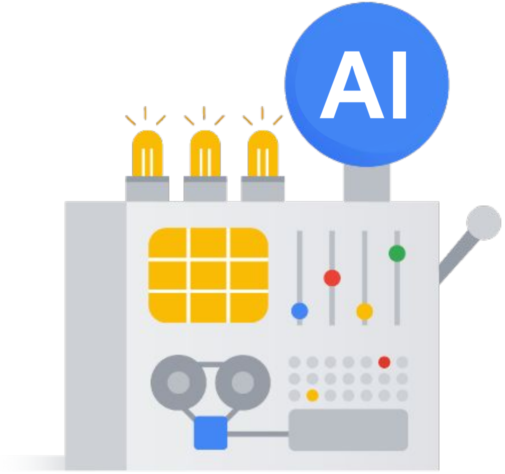
How do you address Interpretability?

- Plan out your options to pursue interpretability
- Treat interpretability as a core part of the UX
- Design the model to be interpretable
- Choose metrics to reflect the end-goal and the end-task
- Understand the trained model
- Communicate explanations to model users
- Test, test, test



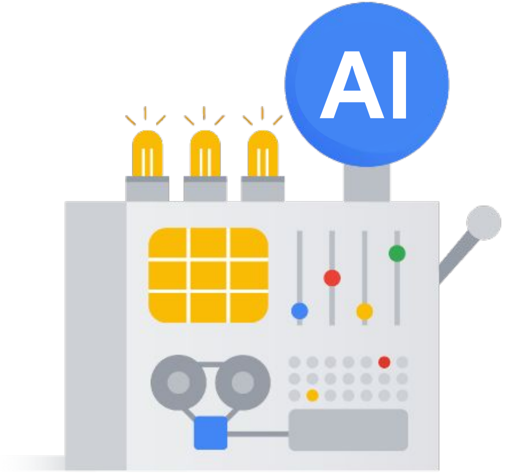
How do you address Interpretability?

- **Plan out your options to pursue interpretability**
- Treat interpretability as a core part of the UX
- Design the model to be interpretable
- Choose metrics to reflect the end-goal and the end-task
- Understand the trained model
- Communicate explanations to model users
- Test, test, test



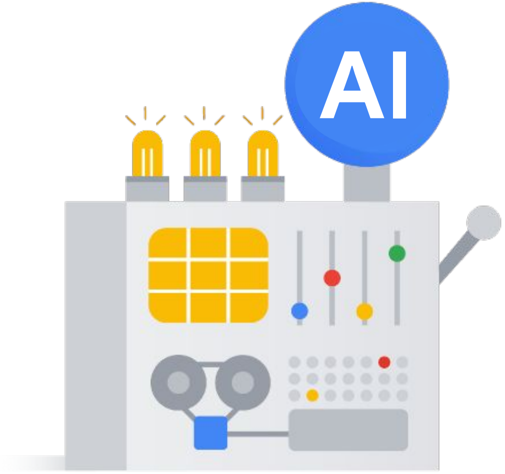
How do you address Interpretability?

- Plan out your options to pursue interpretability
- **Treat interpretability as a core part of the UX**
- Design the model to be interpretable
- Choose metrics to reflect the end-goal and the end-task
- Understand the trained model
- Communicate explanations to model users
- Test, test, test



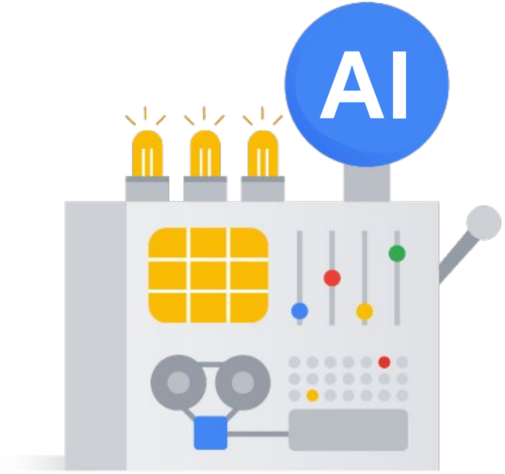
How do you address Interpretability?

- Plan out your options to pursue interpretability
- Treat interpretability as a core part of the UX
- **Design the model to be interpretable**
- Choose metrics to reflect the end-goal and the end-task
- Understand the trained model
- Communicate explanations to model users
- Test, test, test



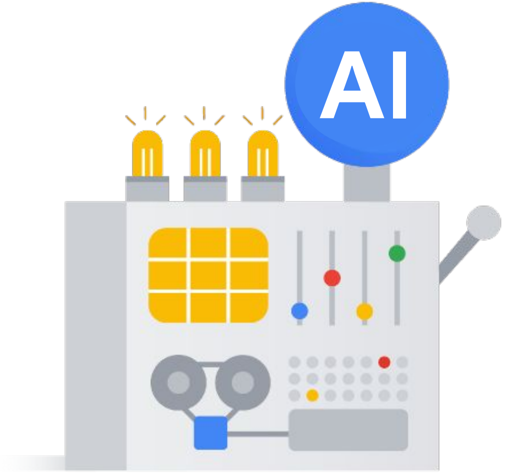
How do you address Interpretability?

- Plan out your options to pursue interpretability
- Treat interpretability as a core part of the UX
- Design the model to be interpretable
- **Choose metrics to reflect the end-goal and the end-task**
- Understand the trained model
- Communicate explanations to model users
- Test, test, test



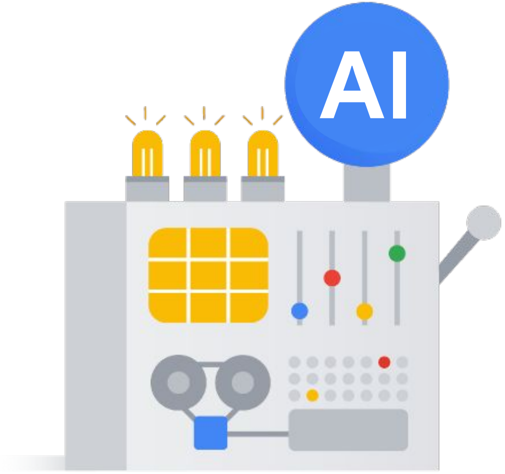
How do you address Interpretability?

- Plan out your options to pursue interpretability
- Treat interpretability as a core part of the UX
- Design the model to be interpretable
- Choose metrics to reflect the end-goal and the end-task
- **Understand the trained model**
- Communicate explanations to model users
- Test, test, test



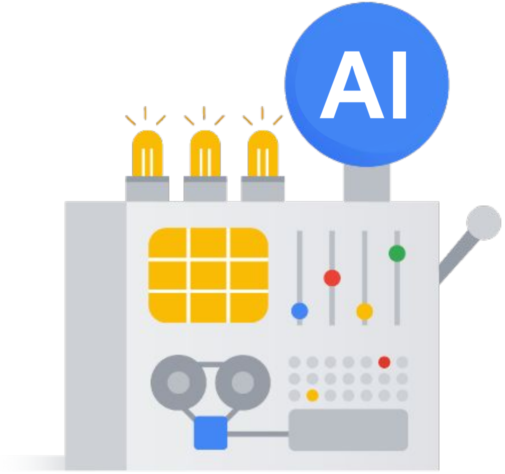
How do you address Interpretability?

- Plan out your options to pursue interpretability
- Treat interpretability as a core part of the UX
- Design the model to be interpretable
- Choose metrics to reflect the end-goal and the end-task
- Understand the trained model
- **Communicate explanations to model users**
- Test, test, test



How do you address Interpretability?

- Plan out your options to pursue interpretability
- Treat interpretability as a core part of the UX
- Design the model to be interpretable
- Choose metrics to reflect the end-goal and the end-task
- Understand the trained model
- Communicate explanations to model users
- **Test, test, test**



Topics

01	Overview of Interpretability
02	Metrics Selection
03	Taxonomy of interpretability in ML Models
04	Tools to Study Interpretability
05	Hands-on Lab

A metric is a statistic that you care about

Technical metric

Business metric

Different metrics are used for different ML scenarios

Technical

Learning Paradigm

- Supervised
- Unsupervised
- Reinforcement Learning

Application

- NLP
- Computer Vision
- GenAI for Text
- ...

Business

Commercial

- Generating revenue
- Reducing costs

Engagement

- Customer engagement and satisfaction
- Brand image

Operational

- Efficiency
- Cost-effectiveness

Different metrics can be applied to different ML models and use cases

Technical
- Learning Paradigm
- Application

Business
- Commercial
- Engagement
- Operational

ML Learning Paradigms

Supervised

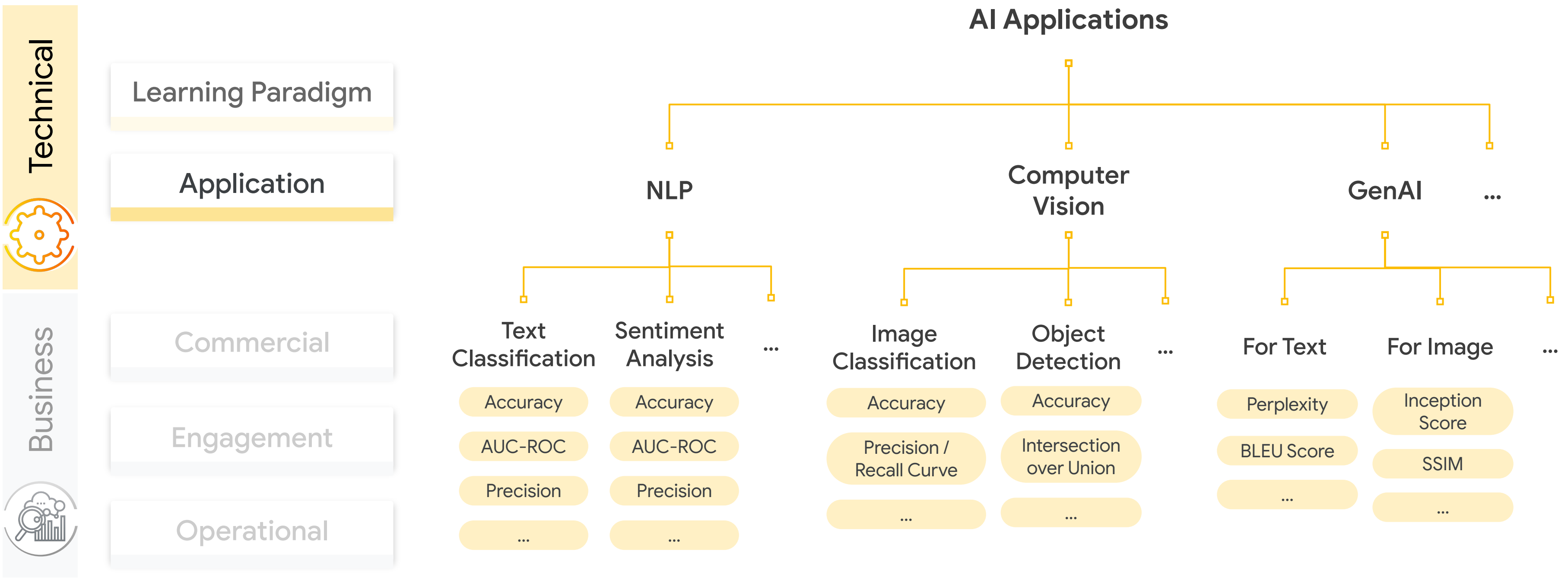
Classification
- Accuracy
- Log Loss
- Precision
- Recall
- ...

Regression
- MAE
- MSE
- RMSE
- R^2
- ...

Unsupervised
- Mutual Information
- Silhouette Coefficient
- Davies-Buldin Index
- ...

Reinforcement Learning
- Dispersion across Time
- Risk across Runs
- ...

Different metrics can be applied to different ML models and use cases



Different metrics can be applied to different ML models and use cases

Technical

Learning Paradigm

Data Type

Application

Business

Commercial	Growth	Sales Conversion Rate	Average Order Value	Customer Acquisition Cost	..
Engagement	Customer Retention Rate	Customer Churn Rate	Net Promoter Score	Customer Lifetime Value	..
Operational	Savings	Resource Utilization	Response Time	Process Time Reduction	..

Different metrics can be applied to different ML models and use cases

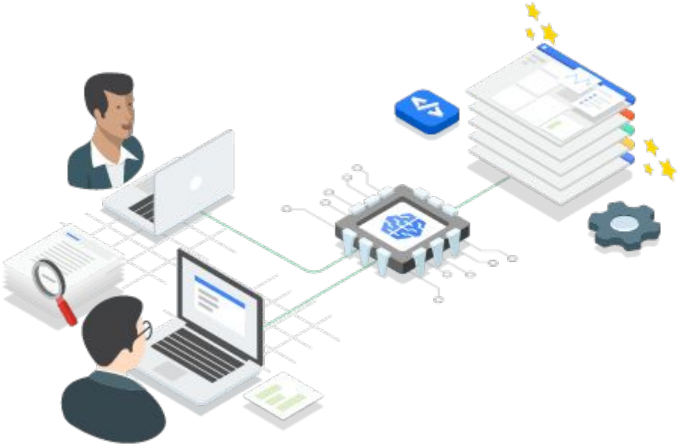
Technical	Learning Paradigm	Accuracy	Precision	AUC ROC	MAE	..
Technical	Application	Log Loss	Recall	Precision / Recall Curve	RMSE	..
Business	Commercial	Growth	Sales Conversion Rate	Average Order Value	Customer Acquisition Cost	..
Business	Engagement	Customer Retention Rate	Customer Churn Rate	Net Promoter Score	Customer Lifetime Value	..
Business	Operational	Savings	Resource Utilization	Response Time	Process Time Reduction	..

Some best practices for metrics selection are:

- Define problem type and goals early-on
- For classification, look at per-class metrics individually if possible
- For regression, evaluate errors proportionally to the label
- Check metrics for important data slices

Some best practices for metrics selection are:

- **Define problem type and goals early-on**
- For classification, look at per-class metrics individually if possible
- For regression, evaluate errors proportionally to the label
- Check metrics for important data slices

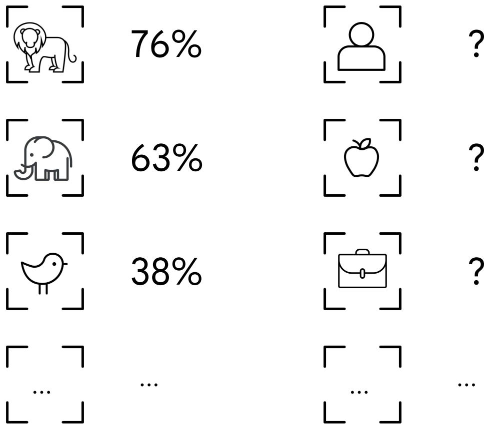


Some best practices for metrics selection are:

- Define problem type and goals early-on
- **For classification, look at per-class metrics individually if possible**
- For regression, evaluate errors proportionally to the label
- Check metrics for important data slices

Object Detection

Overall performance: 70% ACC
Per-class performance:



Some best practices for metrics selection are:

- Define problem type and goals early-on
- For classification, look at per-class metrics individually if possible
- **For regression, evaluate errors proportionally to the label**
- Check metrics for important data slices

Price prediction |

y_pred	*y_true*	*MAE*	***MAPE***
$5	$10	5	100%
$100	$105	5	5%
...	...	...	...

Some best practices for metrics selection are:

- Define problem type and goals early-on
- For classification, look at per-class metrics individually if possible
- For regression, evaluate errors proportionally to the label
- **Check metrics for important data slices**

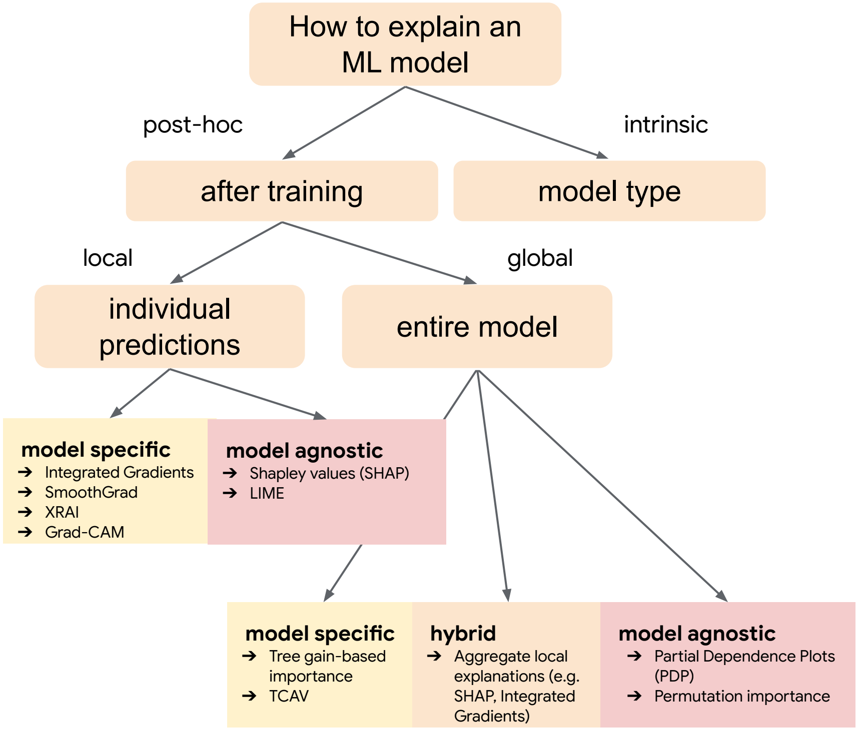
Income prediction

Feature		*Count*
Name	*Values*	
gender	Female	33%
	Male	67%
age	<30	48%
	>=30	52%
...	...	...

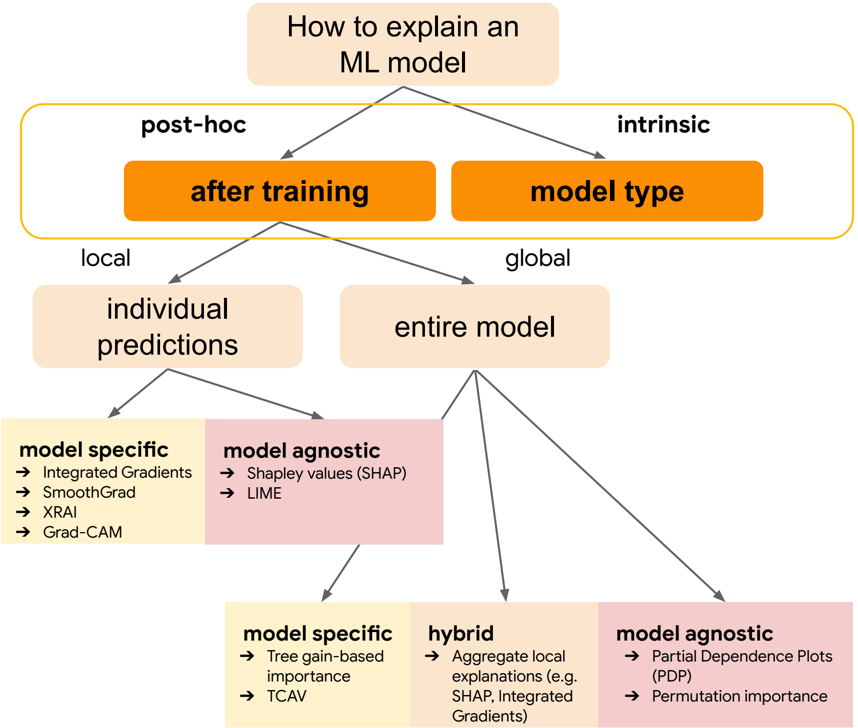
Topics

01	Overview of Interpretability
02	Metric Selection
03	Taxonomy of interpretability in ML Models
04	Tools to Study Interpretability
05	Hands-on Lab

How do you explain an ML model?



How do you explain an ML model?



Post-hoc

Apply post-training methods.

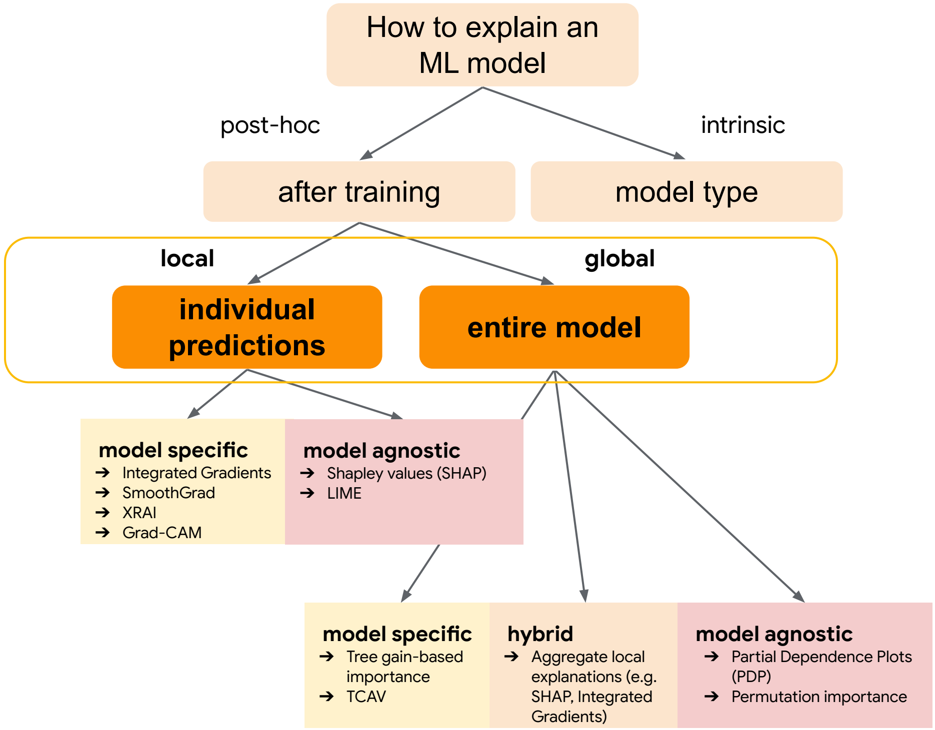
- ✓ For non-intrinsic models
- ✓ For a standardized approach across model types

Intrinsic

Apply specialized methods to the model type.

- ✓ For simple models *(linear models, decision tree, bayesian networks, ...)*

How do you explain an ML model?



Local

Interpretability of individual predictions or a small part of the model's prediction space.

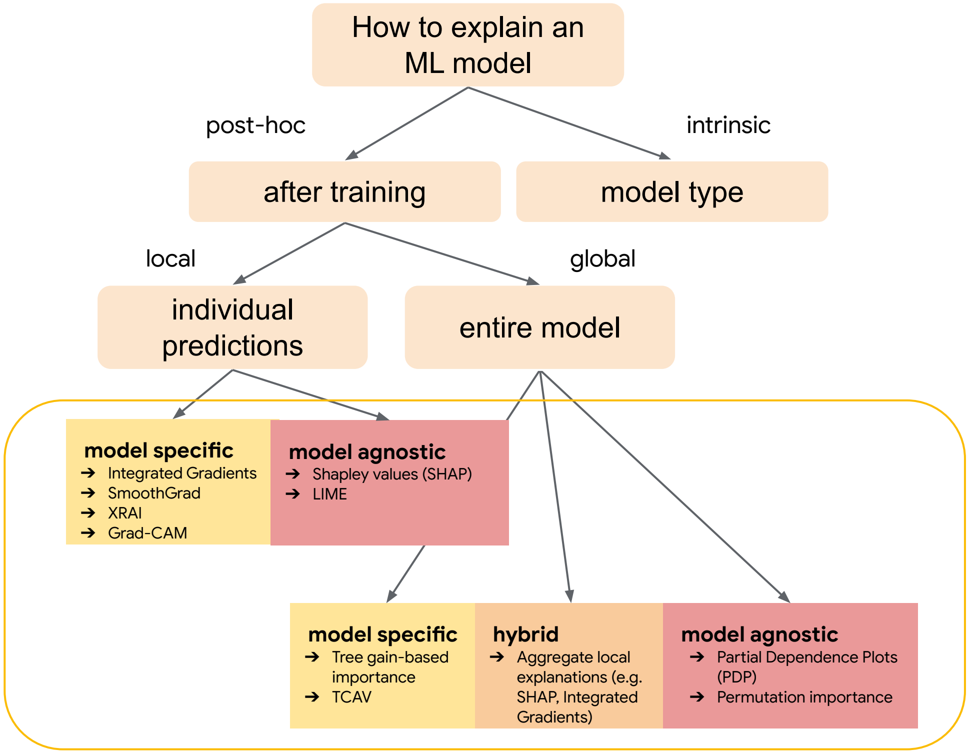
✓ Higher precision
✗ Lower recall

Global

Aggregated, ranked contributions of input variables for the entire model's prediction space.

✓ Higher recall
✗ Lower precision

How do you explain an ML model?



Model specific

Apply specialized post-training methods to the model type.

Model agnostic

Apply generic post-training method for any model.

(Global) Model-agnostic post-hoc methods: Permutation Feature Importance

Measures the importance of a feature by calculating the difference in the model's prediction error after permuting the feature.

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

(Global) Model-agnostic post-hoc methods: Permutation Feature Importance

Measures the importance of a feature by calculating the difference in the model's prediction error after permuting the feature.

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

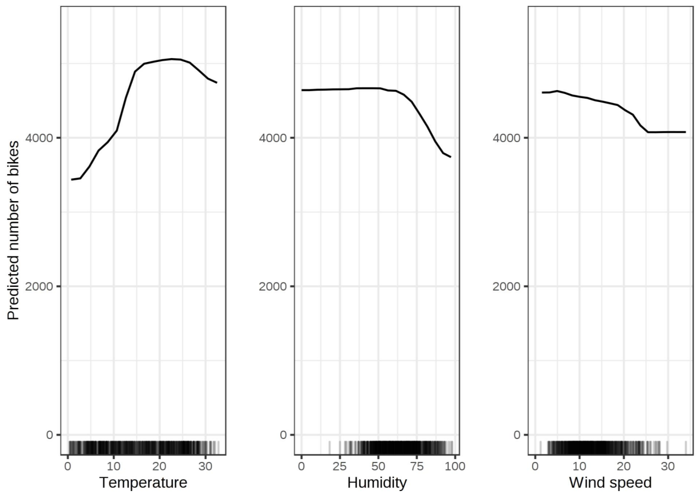
- ✓ Very intuitive
- ✓ Easy to implement
- ✓ Highly compressed global insight
- ✓ No re-training needed
- ✓ All features interactions are accounted for

- ✗ Unreliable for correlated features
- ✗ No insights into individual predictions
- ✗ Needs support of feature distribution view
- ✗ Results can vary with different permutations

This method is **meaningful** and partially **accurate** , but it is **not complete** and **not consistent**.

(Global) Model-agnostic post-hoc methods: Partial Dependence Plots (PDPs)

Shows the marginal effect one or two features have on the predicted outcome of a machine learning model if you force all data points to assume that feature value.

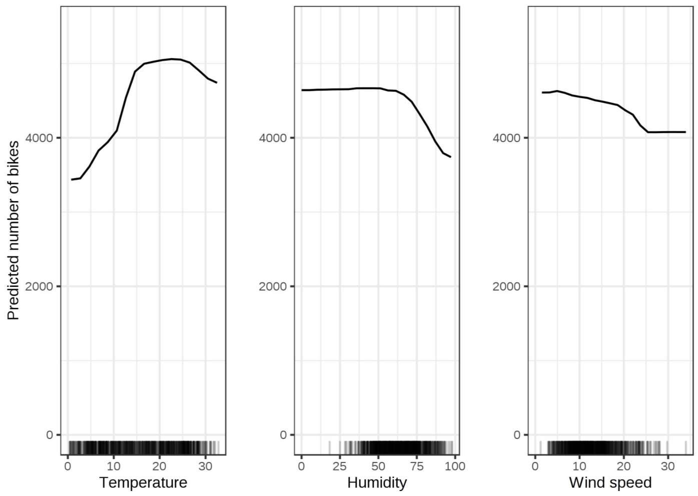


(Global) Model-agnostic post-hoc methods: Partial Dependence Plots (PDPs)

Shows the marginal effect one or two features have on the predicted outcome of a machine learning model if you force all data points to assume that feature value.

- ✓ Very intuitive
- ✓ Easy to implement

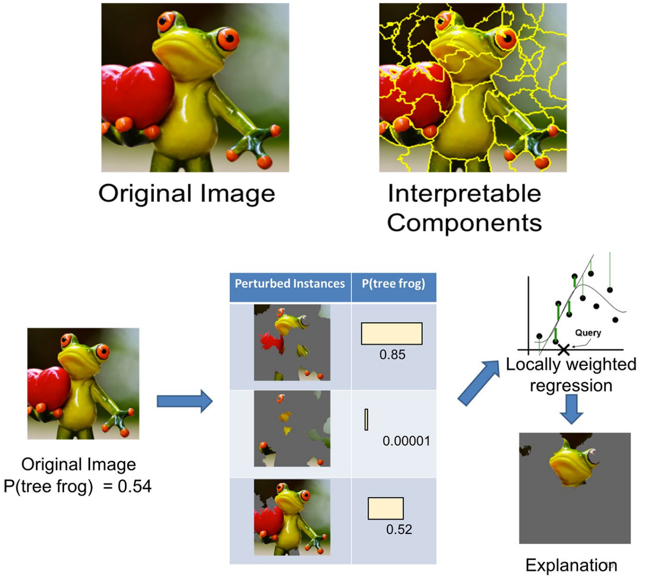
- ✗ Features are assumed to be independent
- ✗ Missing insights for individual predictions
- ✗ At most two features
- ✗ Needs support of feature distribution view



This method is **meaningful**, partially **accurate**, and **consistent**, but it is **not complete**

(Local) Model-agnostic post-hoc methods: LIME

Creates explanations by approximating the underlying model locally with an interpretable one (usually linear or decision tree).



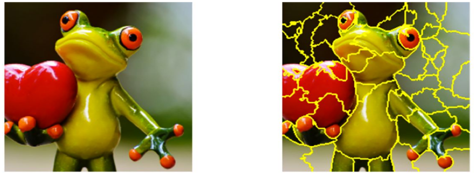
Sources: Marco Tulio Ribeiro, Pixabay

(Local) Model-agnostic post-hoc methods: LIME

Creates explanations by approximating the underlying model locally with an interpretable one (usually linear or decision tree).

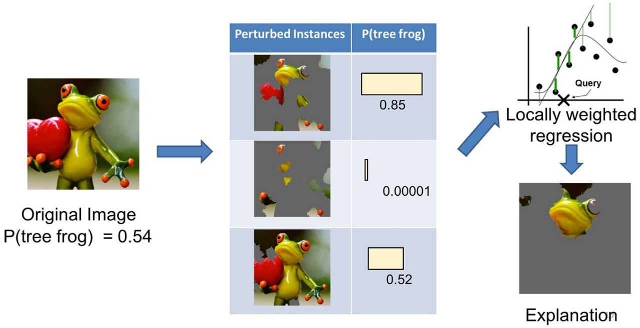
- ✓ Very intuitive
- ✓ Provides interpretations from local models
- ✓ Can work on text, images, and tabular data
- ✓ Can do global interpretation with SP-LIME*

- ✗ Linear assumption reduces accuracy
- ✗ Only works on individual predictions
- ✗ Results can vary upon generation of different synthetic data



Original Image

Interpretable Components

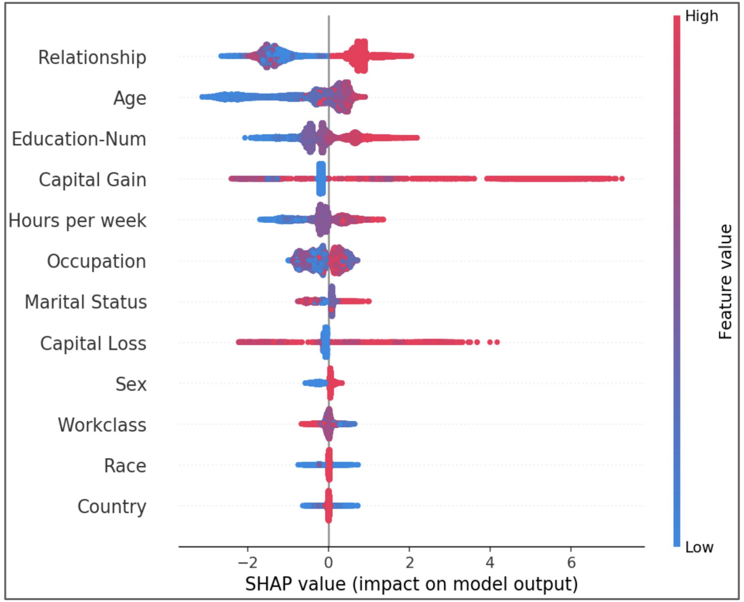


Sources: Marco Tulio Ribeiro, Pixabay

This method is **meaningful**, but it is **not accurate** and **not complete** and **not consistent**.

(Local) Model-agnostic post-hoc methods: SHAP

Generates individual prediction features scores that can be aggregated for global model feature importances on tabular and text data.

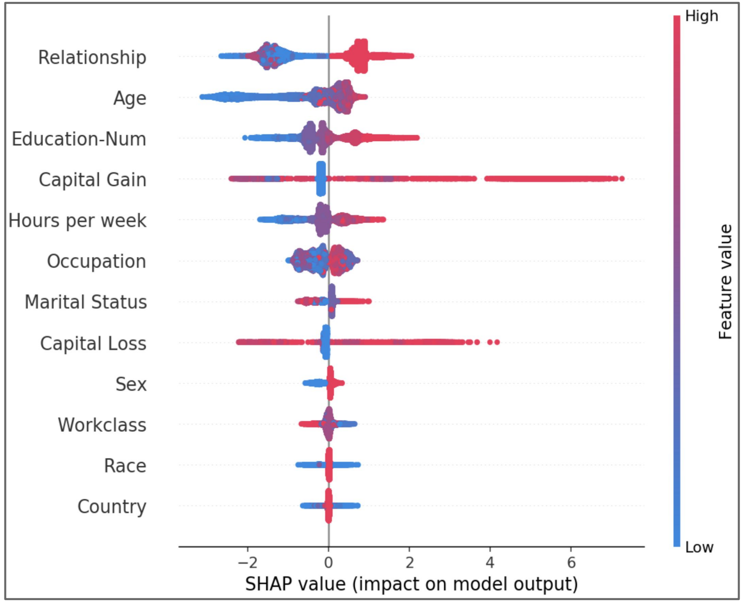


(Local) Model-agnostic post-hoc methods: SHAP

Generates individual prediction features scores that can be aggregated for global model feature importances on tabular and text data.

- ✓ Easy to interpret
- ✓ Compact representation
- ✓ Individual explanation can be aggregated into global model explanations
- ✓ All contributing features are considered

- ✗ Ignores feature interactions
- ✗ Can create new data points that may be irrepresentative
- ✗ Sampling provides an approximate accuracy
- ✗ Computational cost is high on large feature sets



This method is mostly **meaningful**, mostly **accurate**, **complete**, and mostly **consistent**.

Topics

01 Overview of Interpretability

02 Metric Selection

03 Taxonomy of interpretability in ML Models

04 Tools to Study Interpretability

05 Hands-on Lab

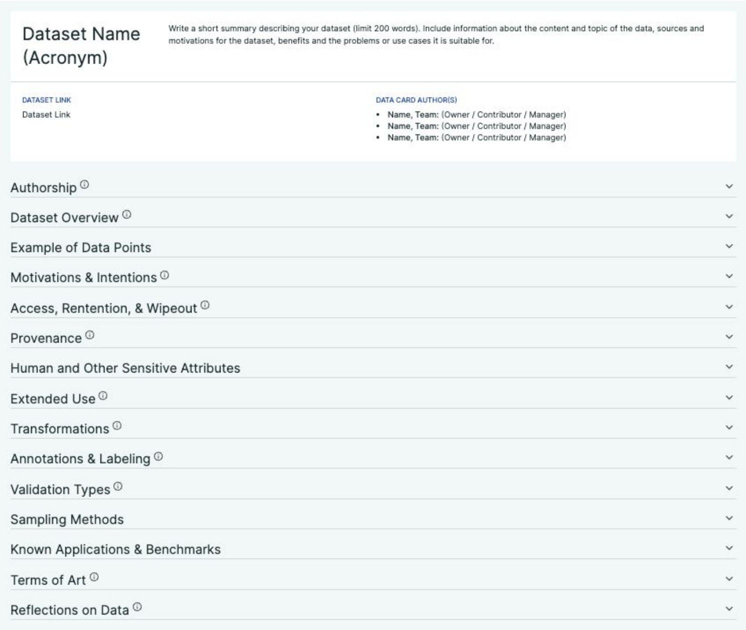
What are good tools to study interpretability?

Data Card Playbook
&
Model Card Toolkit

Learning Interpretability Tool (LIT)

Vertex Explainable AI

Data Card Playbook : a toolkit for transparency in AI dataset documentation


Dataset Name (Acronym)

Write a short summary describing your dataset (limit 200 words). Include information about the content and topic of the data, sources and motivations for the dataset, benefits and the problems or use cases it is suitable for.

DATASET LINK
Dataset Link

DATA CARD AUTHOR(S)
- Name, Team: (Owner / Contributor / Manager)
- Name, Team: (Owner / Contributor / Manager)
- Name, Team: (Owner / Contributor / Manager)

Authorship
Dataset Overview
Example of Data Points
Motivations & Intentions
Access, Rentention, & Wipeout
Provenance
Human and Other Sensitive Attributes
Extended Use
Transformations
Annotations & Labeling
Validation Types
Sampling Methods
Known Applications & Benchmarks
Terms of Art
Reflections on Data

https://sites.research.google/datacardsplaybook/

Model Card Toolkit : a toolkit for transparency in AI model documentation

Model Cards are jinja templates.

A few pre-made templates exist, but you can freely edit them or create your own.

```
import model_card_toolkit

# Initialize the Model Card Toolkit with a path
to store generate assets
model_card_output_path = ...
mct = model_card_toolkit.ModelCardToolkit(
    model_card_output_path
)

# Initialize the model_card_toolkit.ModelCard,
which can be freely populated
model_card = mct.scaffold_assets()
model_card.model_details.name = 'My Model'
model_card.(...)

# Write the model card data to a JSON file
mct.update_model_card_json(model_card)

# Return the model card document as an HTML page
html = mct.export_format()
```

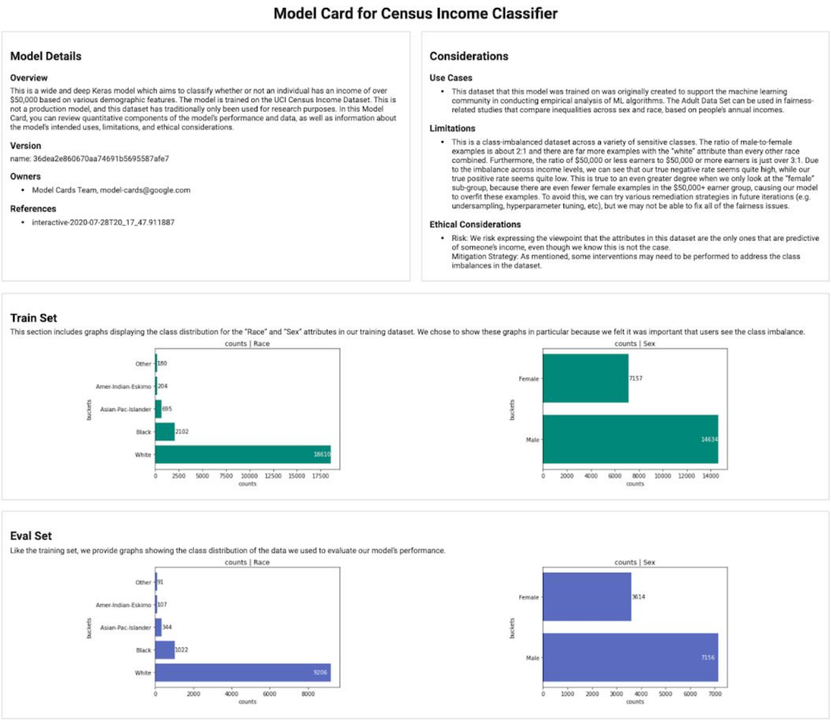
https://www.tensorflow.org/responsible_ai/model_card_toolkit/guide

Model Card Toolkit : a toolkit for transparency in AI model documentation

Improves communication between model builders and product developers.

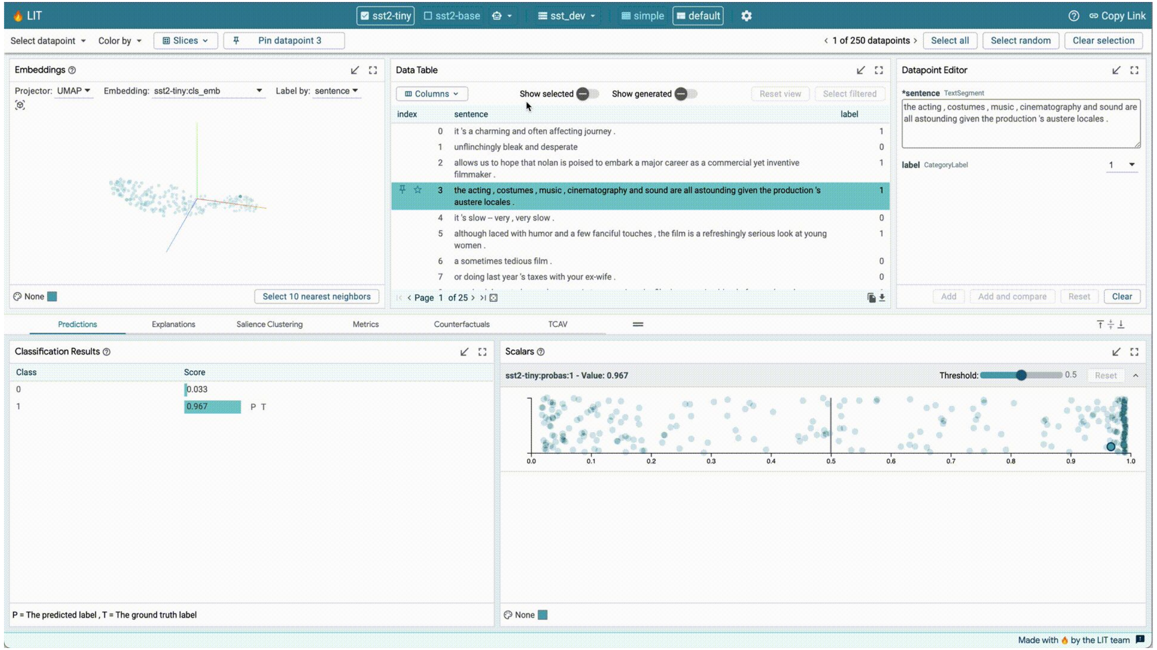
Educates users about ML models.

Provides transparency for public oversight.



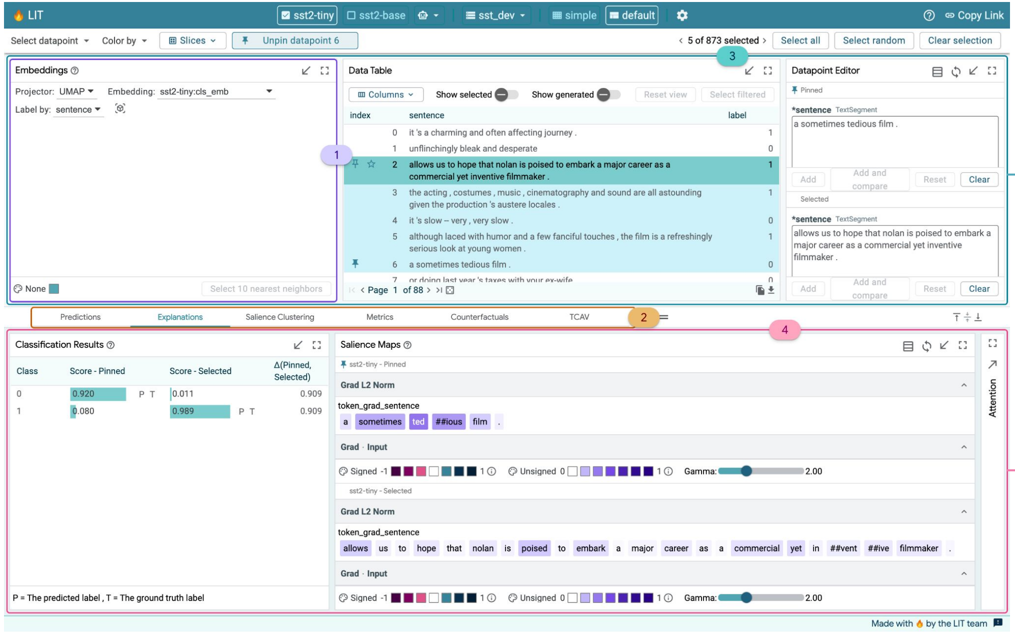
https://www.tensorflow.org/responsible_ai/model_card_toolkit/guide

🔥 Learning Interpretability Tool : an open-source platform for interpretability



https://pair-code.github.io/lit/

🔥 Learning Interpretability Tool : an open-source platform for interpretability

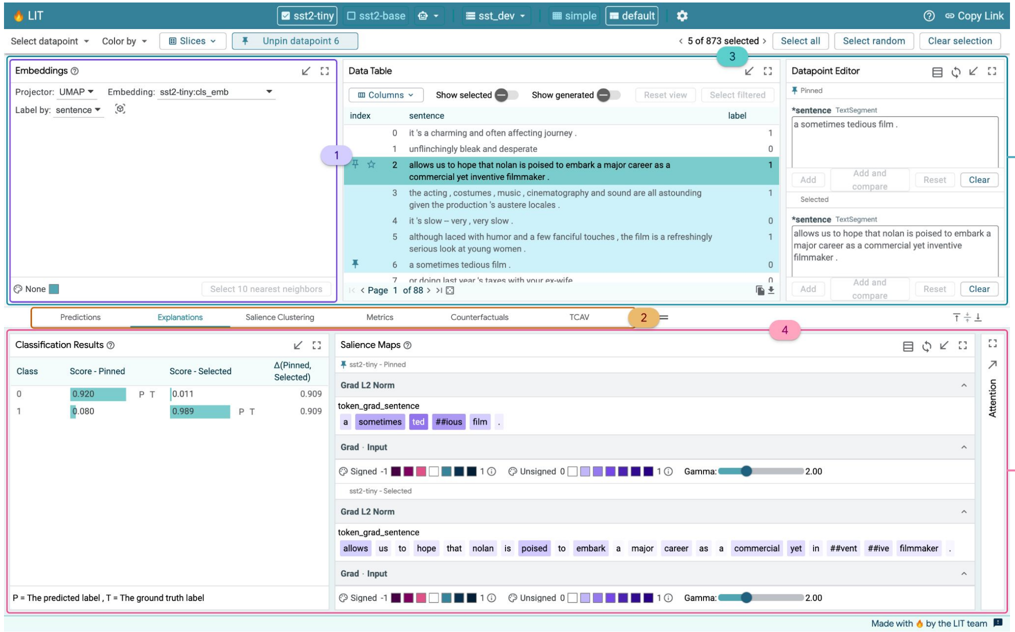


Main workspace

Group-based workspace

https://pair-code.github.io/lit/

🔥 Learning Interpretability Tool : an open-source platform for interpretability



Embeddings
Explore UMAP / TSNE model embeddings.

Data Table
Explore, navigate, and select data.

Datapoint Editor
Deep-dive into individual examples.

Slice Editor
Create and manage data slices.

Performance
Compare models overall or by slices.

Predictions
Explore model results on single data points.

Explanations
Investigate attention for various data points.

Counterfactuals
Autogenerate new data points for evaluation.

https://pair-code.github.io/lit/

Vertex Explainable AI : Google Cloud managed service for interpretability

Example-based explanations

Return a list of examples that are most similar to the input.

Feature-based explanations

Return feature attributions, i.e. contributions, of each feature.

dmlc XGBoost

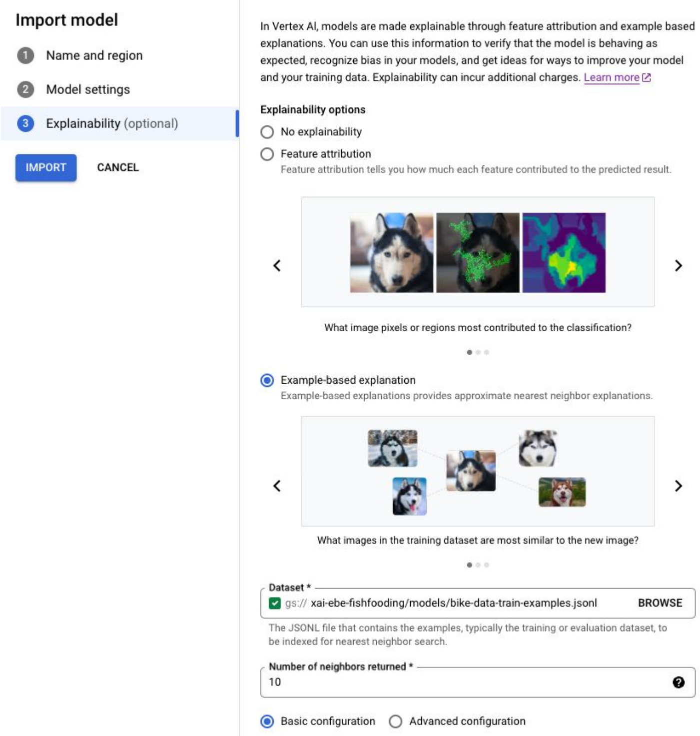
https://cloud.google.com/vertex-ai/docs/explainable-ai

Vertex Explainable AI : Google Cloud managed service for interpretability

Set up explanations for custom models via:

Console gcloud CLI REST Python

Simply import the model in Model Registry, and configure your desired explanations in the Explainability tab!



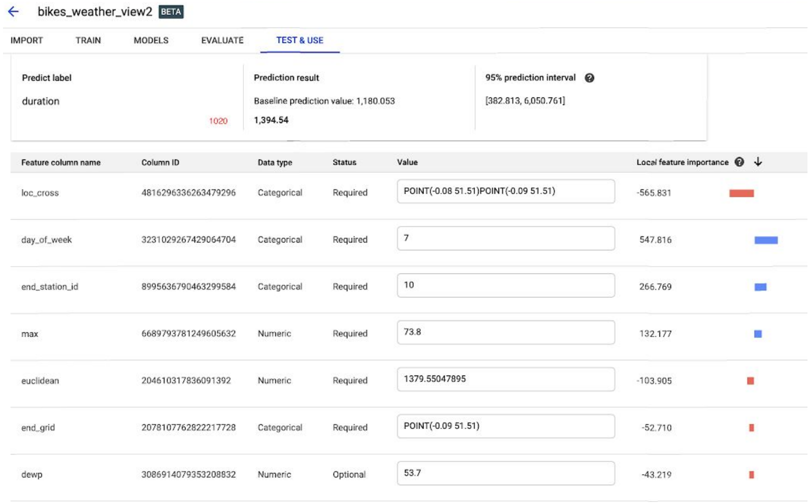
https://cloud.google.com/vertex-ai/docs/explainable-ai

Vertex Explainable AI : Google Cloud managed service for interpretability

BigQuery ML

- AI Explanation functions
  - ML.EXPLAIN_PREDICT
  - ML.EXPLAIN_FORECAST
  - ML.GLOBAL_EXPLAIN
  - ML.FEATURE_IMPORTANCE
  - ML.ADVANCED_WEIGHTS

AutoML



Feature column name	Column ID	Data type	Status	Value	Local feature importance
loc_cross	4816296336263479296	Categorical	Required	POINT(-0.08 51.51)POINT(-0.09 51.51)	-565.831
day_of_week	3231029267429064704	Categorical	Required	7	547.816
end_station_id	8995636790463299584	Categorical	Required	10	266.769
max	6689793781249605632	Numeric	Required	73.8	132.177
euclidean	204610317836091392	Numeric	Required	1379.55047895	-103.905
end_grid	2078107762822217728	Categorical	Required	POINT(-0.09 51.51)	-52.710
dewp	3086914079353208832	Numeric	Optional	53.7	-43.219

https://cloud.google.com/vertex-ai/docs/explainable-ai

Topics

01	Overview of Interpretability
02	Metric Selection
03	Taxonomy of interpretability in ML Models
04	Tools to Study Interpretability
05	Hands-on Lab

Lab:
Explaining Text Classification with Vertex Explainable AI