

Interpretable Machine Learning

Giacomo Marangoni

PRISMA Summer School - June 15, 2026

Machine Learning

Classification

- ...

Regression

- ...

Classification

- **Goal** Assign input data to one of a predefined set of categories or classes.
- **Output** Discrete labels (e.g., spam vs. not spam, cat vs. dog, disease present vs. absent).
- **Example** Classifying emails as spam or not spam.

Regression

- **Goal** Predict a continuous value or outcome based on input features.
- **Output** Numeric values or probabilities.
- **Example** Predicting house prices based on features like size, location, and number of bedrooms.

- Data

- Raw information collected from various sources, used to train and evaluate models.
- **Example:** A listing of houses with descriptions.

- Features

- Measurable properties or characteristics extracted from the raw data, used as input to the machine learning model.
- **Example:** Square meters, number of bedrooms, age of the house, location.

- Algorithms

- Computational methods used to build machine learning models by learning patterns from data
- **Example:** Linear Regression, Random Forest Regression

1. Define the problem

- Specify inputs, outputs, target metric, and constraints.

2. Collect and inspect data

- Check data quality, missing values, imbalance, and leakage risks.

3. Split the dataset

- Train set: fit the model.
- Validation set: tune hyperparameters.
- Test set: final unbiased evaluation.

4. Preprocess and engineer features

- Normalize, encode categorical variables, handle missing values.

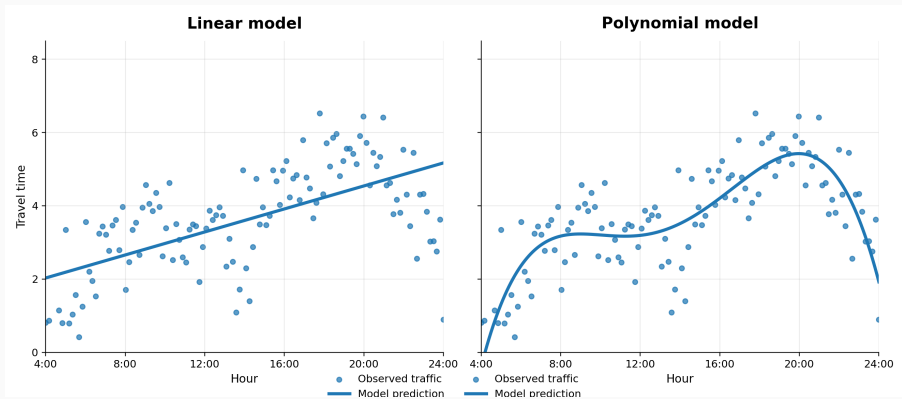
5. Train and tune models

- Compare baselines and candidate models.

6. Evaluate

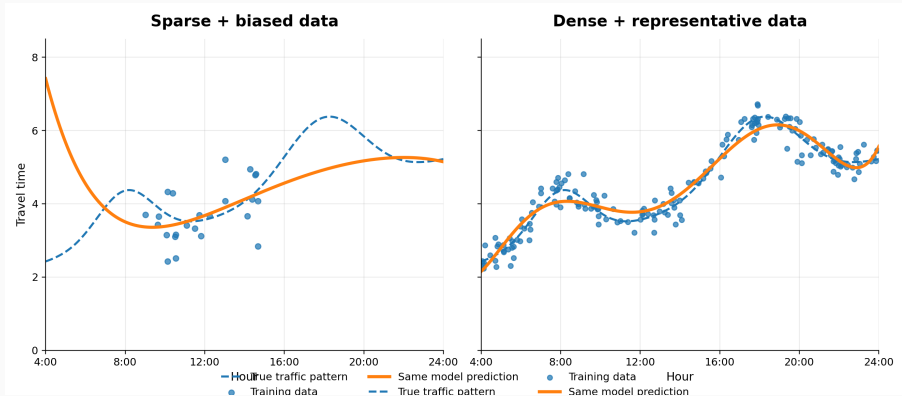
- Report test performance, analyze errors.

Model matters

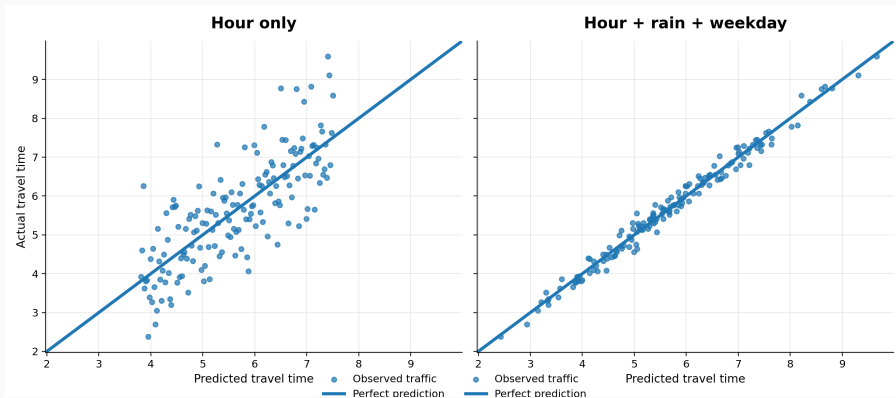


Common ML models

Model	Intuition	Pros	Cons
Linear / Logistic Regression	Fit a straight-line relationship or decision boundary	Fast, interpretable	Limited nonlinear modeling
Decision Tree	Split data using a sequence of if-then rules	Easy to explain and visualize	Overfits easily
Random Forest	Combine predictions from many trees	Robust, strong baseline	Less interpretable
Gradient Boosting / XGBoost	Sequentially fix mistakes of previous trees	Excellent predictive performance	More tuning required
SVM	Find the boundary with the widest gap between classes	Effective on small/medium datasets	Slower on large datasets
k-NN	Predict from the labels of the most similar examples	Simple, no training phase	Slow inference, sensitive to scaling
Naive Bayes	Combine evidence from features assuming independence	Very fast, works with little data	Strong assumptions
Neural Network	Learn useful representations directly from data	Highly flexible and powerful	Data-hungry, less interpretable



Features matter



Predicts a **continuous value** (e.g. temperature)

Common metrics

MAE average absolute error

MSE average squared error

RMSE square root of MSE

R^2 explained variance

Predicts a **discrete class** (e.g. spam/not spam)

Common metrics

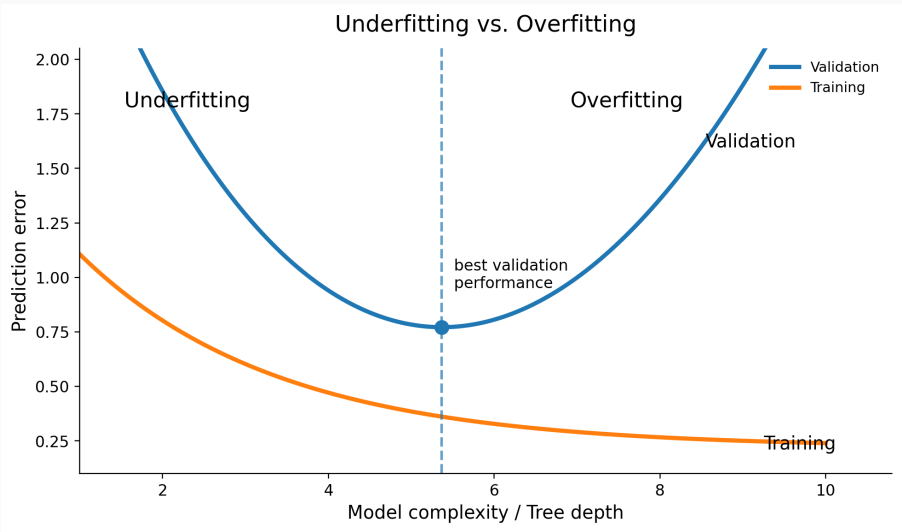
Accuracy fraction correct

Precision reliability of positive predictions

Recall ability to find positives

F1-score balance of precision and recall

Overfitting & Underfitting



High accuracy

low recall

Usually caused by
imbalanced data.

The model mostly predicts the majority class: it looks good overall, but misses many positives.

High precision

low recall

Usually a **conservative model.**

Positive predictions are usually correct, but many true positives are missed.

High recall

low precision

Usually an **aggressive model.**

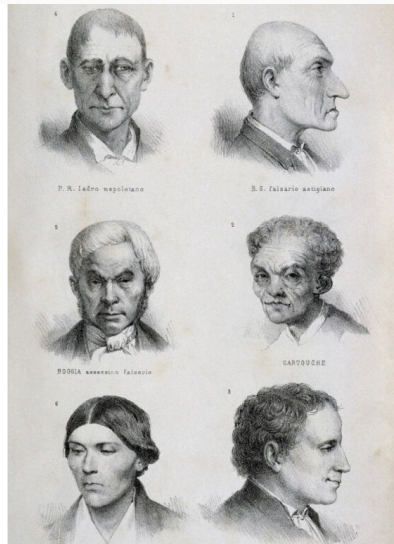
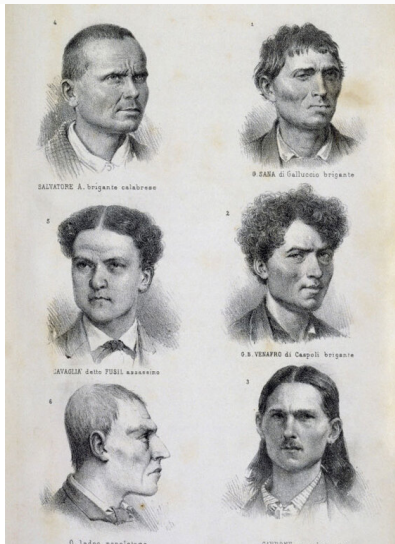
Most positives are found, but many false alarms are created.

Constraints

Constraint	Meaning
Time	Prediction must be fast enough.
Memory	Model must fit on available hardware.
Cost	Training or serving should not be too expensive.
Interpretability	Humans must understand why the model predicts something.
Fairness	Model should not discriminate across groups.
Privacy	Data must be handled securely.
Regulation	Must satisfy legal or industry rules.
Data availability	You may only have limited or noisy data.

Explainable AI

Criminal faces



1 Cesare Lombroso. L'Uomo delinquente (Criminal Man), 1878.

Criminal facial recognition

- Wu and Zhang gathered over 1,800 photos of Chinese men, aged 18-55, without facial hair, scars, or tattoos.
- Approximately 1,100 were of noncriminals, sourced from the internet, including professional networking sites and company staff pages.
- The remaining 700+ were convicted criminals, with photos obtained from police IDs.
- They claimed that their AI algorithm could detect criminal intentions from facial features.



(a) Three samples in criminal ID photo set S_c .

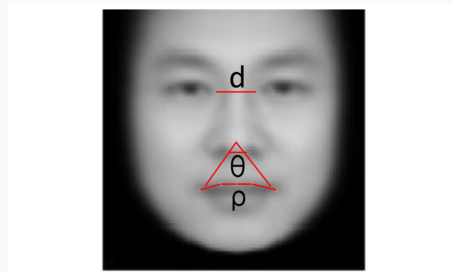


(b) Three samples in non-criminal ID photo set S_n .

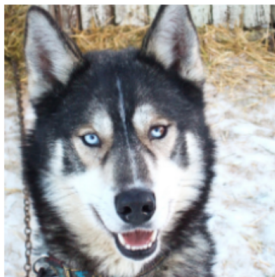
2 Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. arXiv Preprint arXiv:1611.04135, 4038–4052.

What are the issues with this research?

- Noncriminals vs criminals photos
- Convicted vs committing criminals
- Facial features vs expressions



3 Wu & Zhang, discriminative features.



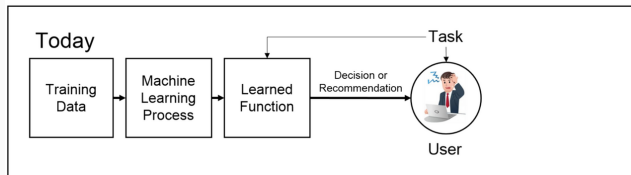
(a) Husky classified as wolf



(b) Explanation

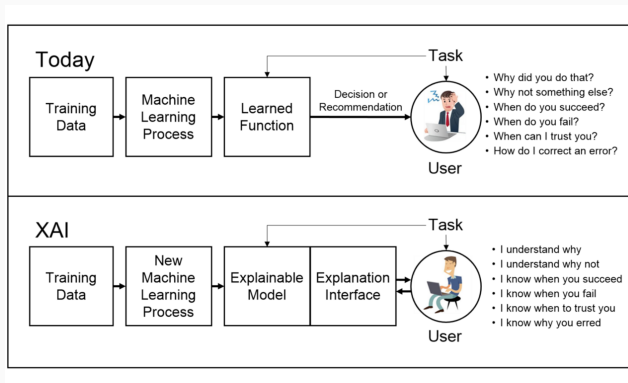
Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27



You are a radiologist testing an AI-powered radiology tool that analyzes chest X-rays to identify potential cases of pneumonia. You prioritize patient cases based on the AI's recommendations. What questions would you like to know the answer of?

Explainable AI



5 Gunning, D. (2016). Broad Agency Announcement - Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), Tech. Rep.

Debug & improve the model: leakage, instability, bias, privacy risks, and poor transferability.

Build trustworthiness: support audit, accessibility, interactivity, fairness checks, and human oversight.

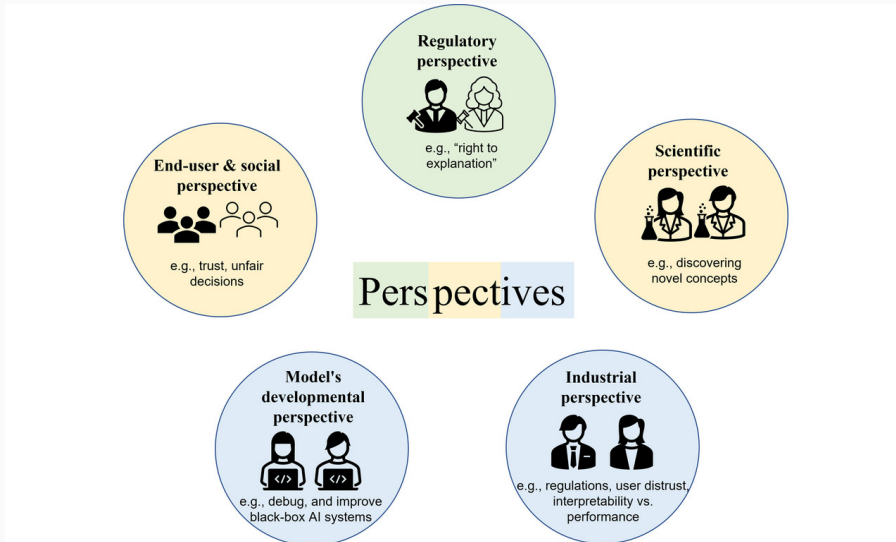
Extract information: understand feature effects, learned relationships, reusable patterns, and possible causal hypotheses.

Caution

Explanations do not prove causality and may also pose further privacy risks.

Source: adapted from Molnar, *Interpretable Machine Learning*, Chs. 2–3.

Actors and perspectives involved



Interpretable vs. explainable ML

Interpretable

Map abstract model behavior into a human-understandable form.

Main question:

What has the model learned?

Typical focus: Model structure, feature effects, learned relationships

Examples:

linear/logistic models, shallow trees, rules, GAMs/EBMs.

Explainable

Interpretation plus context for a user, task, and decision.

Main question:

Why this prediction, for this person, in this situation?

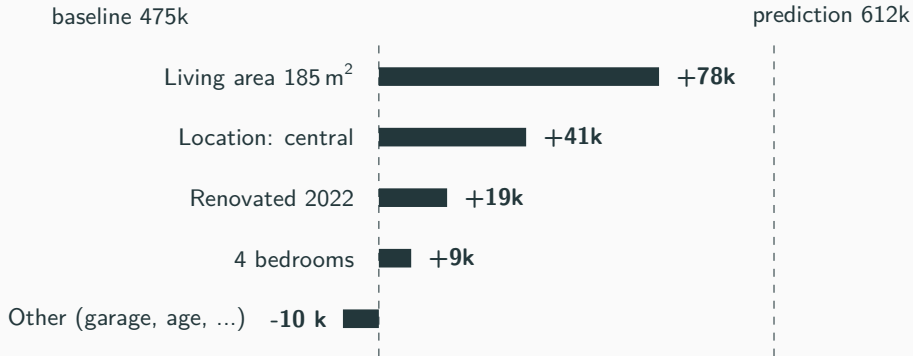
Typical focus: Local reasons, contrastive explanations, justification

Examples:

counterfactuals, local SHAP/LIME, recourse, explanation reports.

Source: Molnar, *Interpretable Machine Learning*, Chapters 2–3.

Explanations: Why does this model value this house at €612k?



Contrast: a suburban twin would be $\approx 571\text{k}$ — location drives most of the gap. **Lever:** only renovation is changeable (-19k if un-renovated).

Flag: 185 m² is in the top 5% of the data — the area effect is extrapolated; trust the $\pm 24\text{k}$ band with caution.

Does the explanation faithfully describe the model?

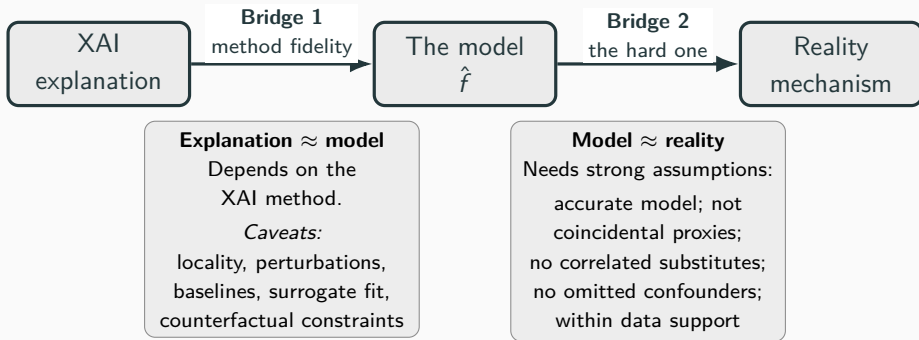
- **Fidelity**: does the explanation faithfully reproduce what the black-box model actually predicts?
- **Stability**: do similar inputs receive similar explanations?
- **Consistency**: do similarly performing models give similar explanations?
- **Certainty**: does the explanation convey the model's confidence in the prediction?
- **Novelty**: does it flag inputs far from the training data, where the explanation is unreliable?

Does it work for a person?

- **Comprehensibility:** can the intended audience (not just an ML expert) understand the explanation?
- **Degree of importance:** are the main reasons clearly ranked?
- **Representativeness:** does it explain one case (local) or many cases (global)?
- **Contrastiveness:** why this outcome instead of another?
- **Selectiveness:** does it focus on the few most relevant causes?
- **Social fit:** is the content tuned to the audience and context?
- **Actionability:** does it suggest what could be changed?

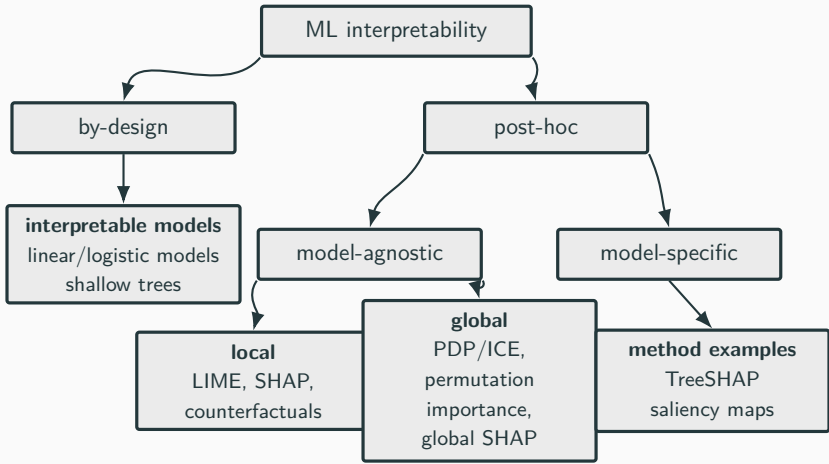
Does it correspond to reality?

- **Accuracy:** does the underlying model predict unseen data well?
- **Non-spuriousness:** does the explanation credit features that have genuine causal relationships to the outcome, and not merely because they correlated with one that does?
- **Consistency with domain knowledge:** does it cohere with established theory, or flag where the model contradicts it?



XAI $\approx$ model: depends on fidelity assumptions • model $\approx$ reality: rarely provable from fit alone

A taxonomy of ML interpretability



Adapted from <https://christophm.github.io/interpretable-ml-book>

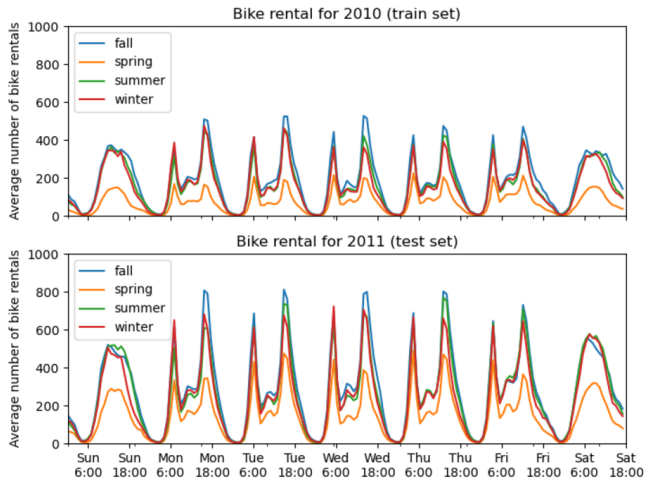
Partial Dependence Plots and Individual Conditional Expectation

Dataset Bike Sharing Rentals from Capital Bikeshare system in Washington, D.C. in 2010/2011

- provides number of bikes rented on an hourly and daily basis
- includes various weather-related features such as temperature, humidity, wind speed

Goal predict the number of bike rentals using weather and season data as well as the datetime information.

Bike Sharing Dataset

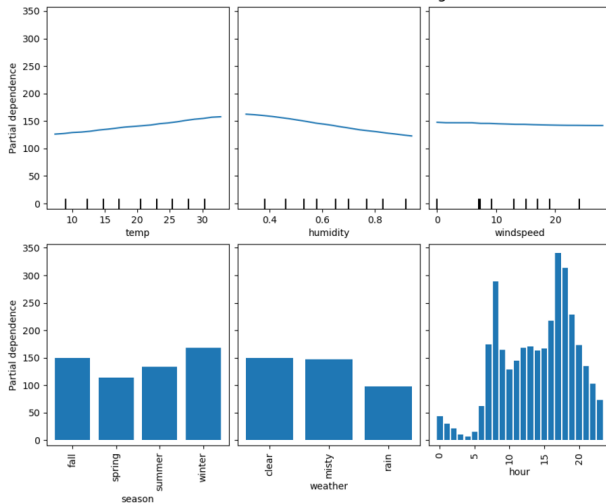


Partial Dependence Plots

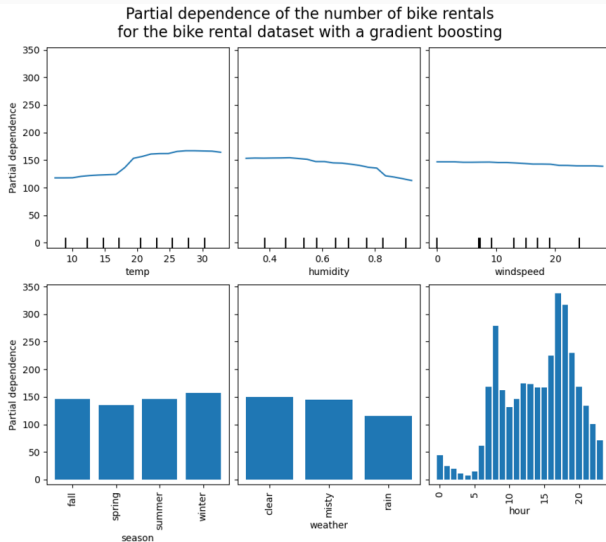
- Select the feature of interest.
- Define a set of feature values to evaluate.
 - Use either all unique values or a grid of points over the observed range (e.g., uniform sampling).
- For each value in the grid:
 - Make a copy of the dataset
 - Replace the selected feature's value for every observation with this grid value.
 - Compute the model's predictions for this modified dataset.
- Average the predictions over all observations.
 - This provides the marginal effect of the feature on the prediction across the distribution of all other features.
- Plot the averaged predictions against the feature values.
 - This visualizes the relationship between a feature and the predicted outcome, while marginalizing over other features to isolate the effect of the target feature.

Partial Dependence Plots

Partial dependence of the number of bike rentals
for the bike rental dataset with an MLPRegressor

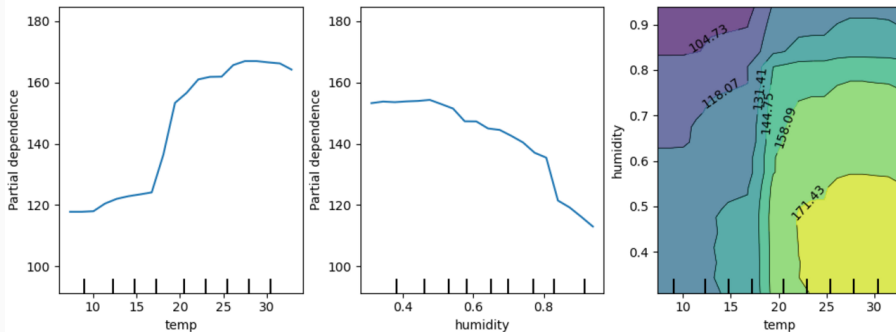


Partial Dependence Plots

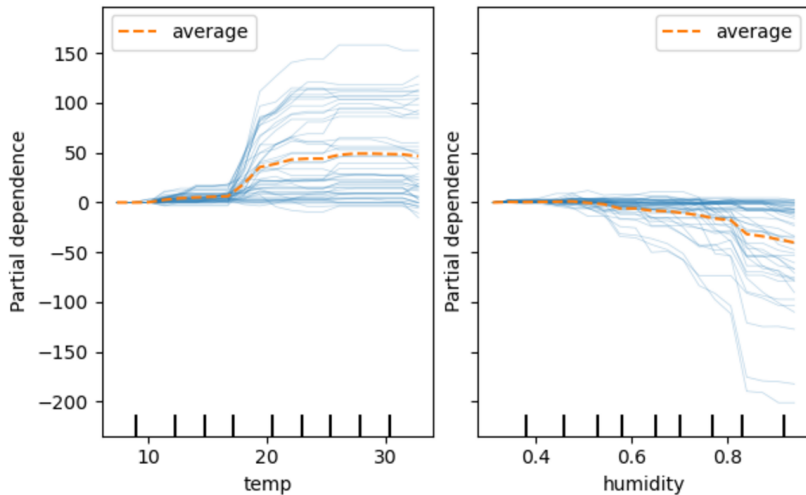


Partial Dependence Plots - 2D

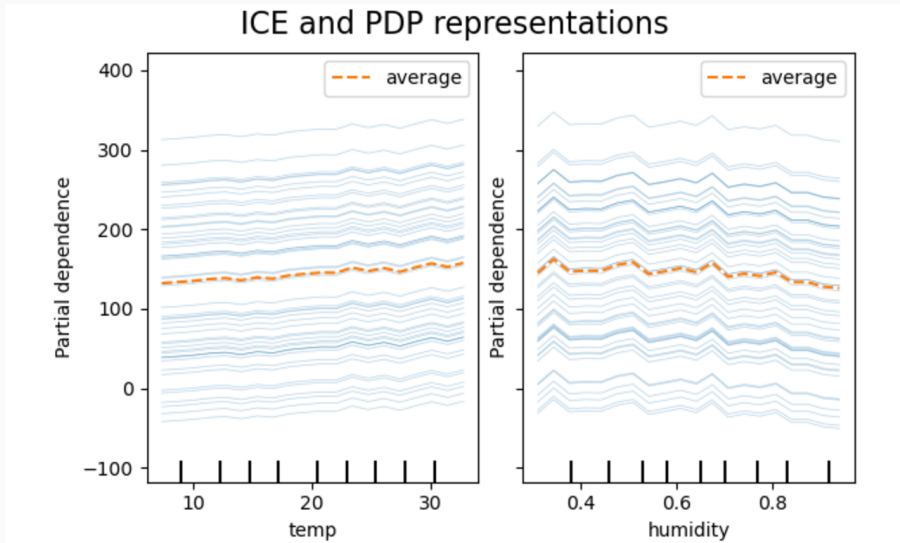
1-way vs 2-way of numerical PDP using gradient boosting



ICE and PDP representations

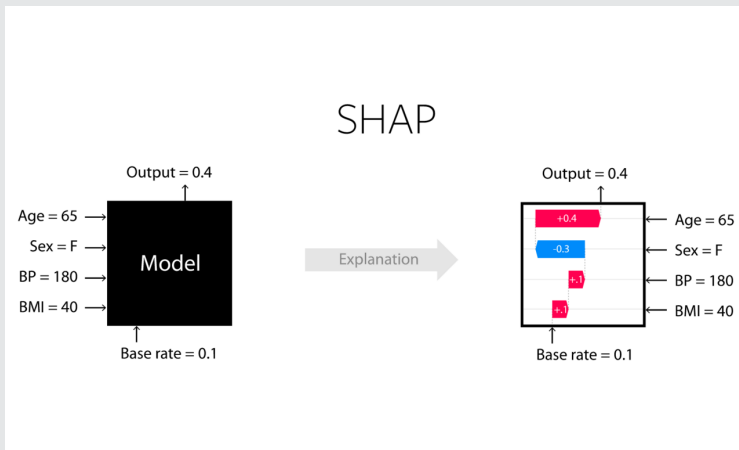


What if model has no interaction between features?



SHAP

SHAP (SHapley Additive exPlanations)





14 Lloyd Shapley, 2012 - Wikipedia

- Imagine a global initiative where \$100 millions reward is given to a set of countries who reduce their overall emissions by 1 million tons via renewables and energy savings.
- How to fairly distribute the reward among participating countries?

Let's say we have three countries involved in this initiative: Country A, Country B, and Country C. Their individual and combined efforts contribute differently to the total emission reduction.

- **Individually:** Country A reduces 200,000 tons, Country B reduces 300,000 tons, and Country C reduces 500,000 tons.
- **In pairs:** A and B together reduce 600,000 tons, A and C reduce 800,000 tons, and B and C reduce 900,000 tons.
- **All together:** When all three countries collaborate, they achieve the goal of reducing 1 million tons.

The marginal contribution

What is the additional value or impact that each country brings to the collective effort, above and beyond what the other countries achieve without it?

If Country A reduces 200,000 tons of emissions on its own, but when it collaborates with Country B, they together reduce 600,000 tons, the marginal contribution of Country A in the coalition with B is not just its individual 200,000 tons, but rather how much more is achieved together than the sum of their individual efforts

By rewarding marginal contributions, the initiative

1. ensures that the distribution is fair and reflects the actual value added by each country's efforts
2. encourages countries to work together rather than compete, leading to more innovative and effective solutions than what would be possible through individual efforts alone.

The marginal contribution

For Country A: Calculate the marginal contribution of Country A in all possible combinations.

- A joining B: $600,000 - 300,000 = 300,000$
- A joining C: $800,000 - 500,000 = 300,000$
- A joining B and C: $1,000,000 - 900,000 = 100,000$
- A working alone: 200,000

The Shapley value for Country A is the average of these contributions.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

- S is a subset of N not including player i .
- $|S|$ is the number of players in coalition S .
- $|N|$ is the total number of players.
- $v(S \cup \{i\})$ is the value of the coalition including player i .
- $v(S)$ is the value of the coalition without player i .
- The summation is over all subsets S of N that do not include player i .

Properties of Shapley values

Efficiency The total payoff distributed among all players equals the total value generated by the grand coalition (the coalition of all players).

Symmetry If two players contribute equally to any coalition they are part of, they receive the same payoff.

Dummy player A player who adds no additional value to any coalition (a "dummy" player) receives a payoff of zero.

Additivity For any two games with payoff functions v and w , the Shapley value of the sum of the games is the sum of the Shapley values of each game.

Cooperative game theory	XAI
Players	Features
Total payoff	Prediction Output
Marginal contributions	How much features change the prediction
Fair distribution	Fair assessment of features' influence

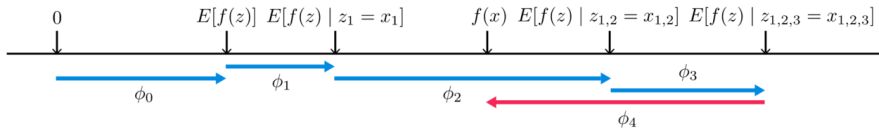
SHAP (SHapley Additive exPlanations)

- The expected prediction if no features are known is $E[f(z)]$.
- The SHAP value is calculated as the change in the expected model prediction when conditioning on that feature. This involves comparing the prediction with and without the feature, averaged over all combinations of other features.

Computational complexity

- Exact computation of SHAP values is often challenging.
- Feature independence and linearity assumptions simplify SHAP computations.
- Methods like Kernel SHAP or Shapley Sampling Values provide practical ways to approximate SHAP values for complex machine learning models.

SHAP (SHapley Additive exPlanations)



```
X100 = shap.utils.sample(X, 100)
```

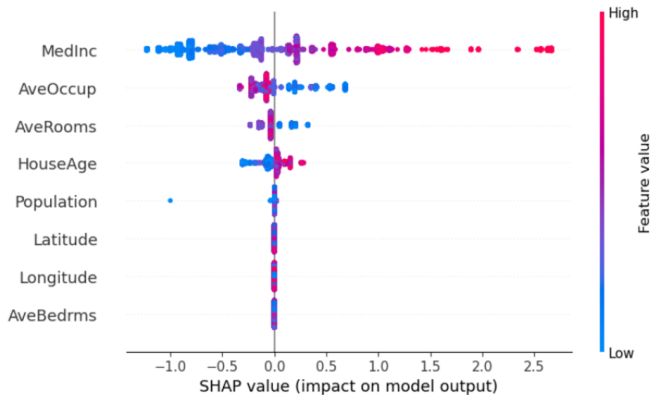
- Provides a representative sample of the dataset.
- Helps comparing impact of features of interest with a baseline.
- Helps isolating the effect of a feature by holding other features constant at their typical values as represented in the background dataset.

Summary plot

```
[27]: explainer = shap.Explainer(model_regression)
      shap_values = explainer(X_test)
      shap_values.shape

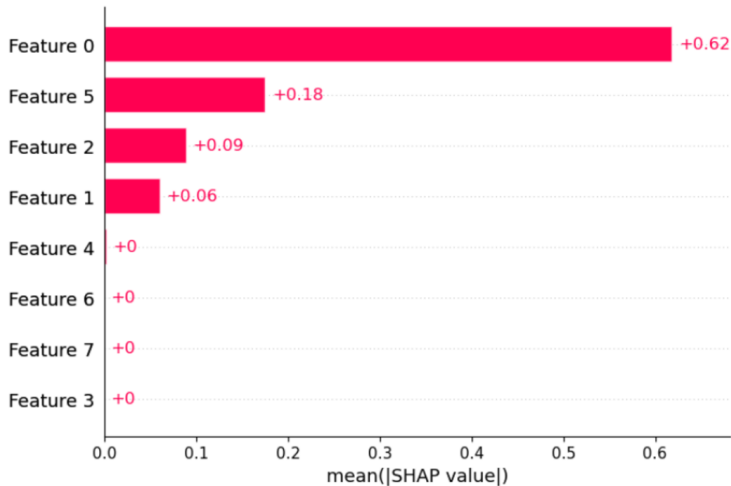
[27]: (4128, 8)

[31]: shap.summary_plot(shap_values, feature_names=feature_names)
```



Bar plot

```
[35]: shap.plots.bar(shap_values)
```



Waterfall plot

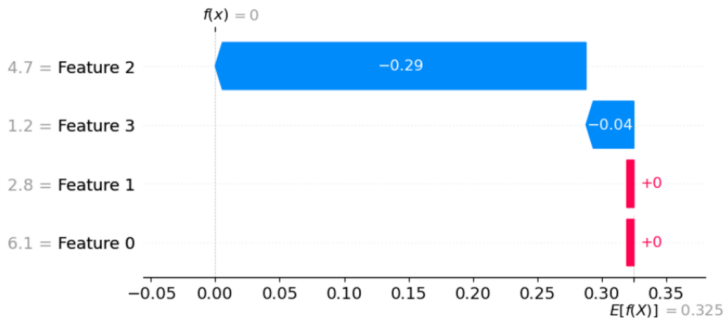
```
•[12]: explainer = shap.Explainer(model_classification)
      shap_values = explainer(X_test)
      shap_values.shape
```

```
[12]: (30, 4, 3)
```

```
[13]: X_test.shape
```

```
[13]: (30, 4)
```

```
[23]: shap.plots.waterfall(shap_values[0,:,2])
```

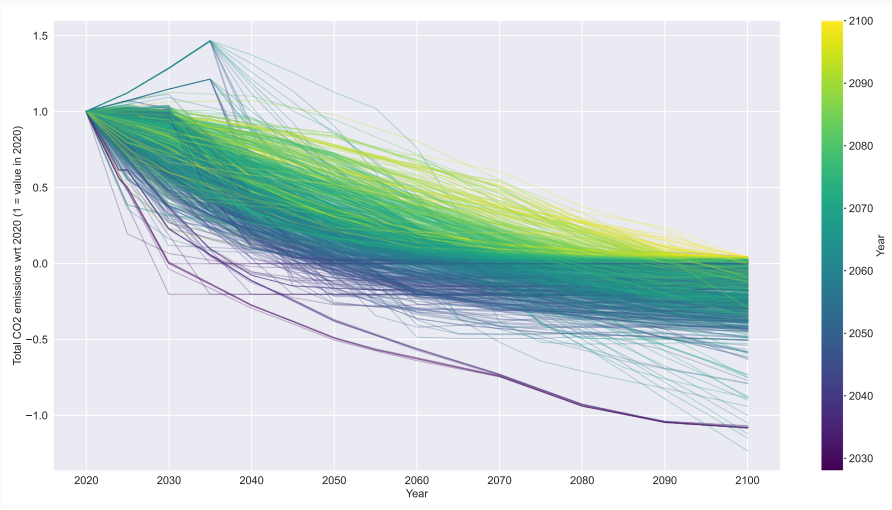


Application to IAMs scenarios study

- Thousands of scenarios are available in the AR6 database.
- Many pathways can reach net-zero global CO2 emissions.
- Some pathways appear more feasible than others.

How can explainable AI help us understand why some net-zero pathways appear more feasible than others, depending on their timing, technology choices, and socioeconomic context?

The pathways to net-zero

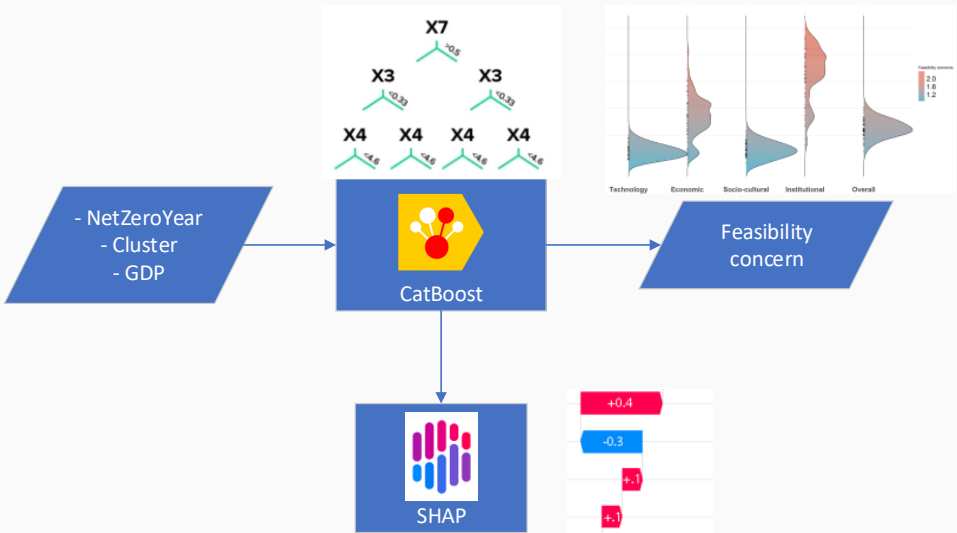


16 Marangoni et al. (in preparation)

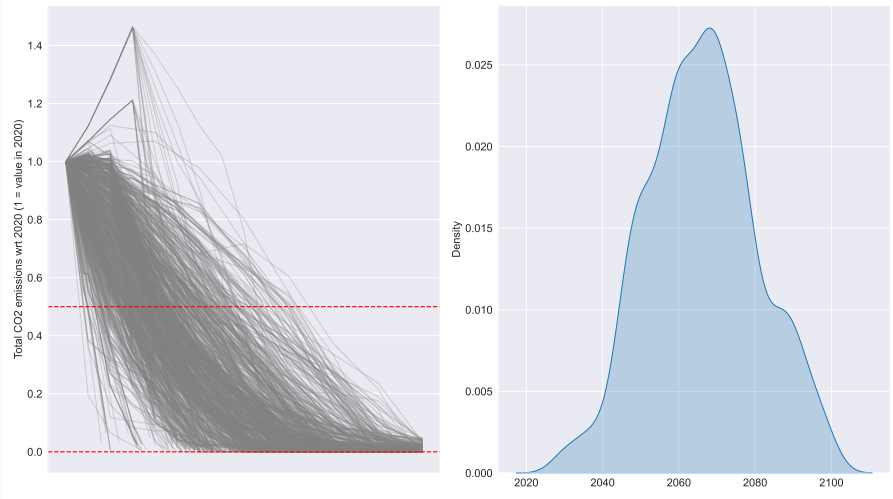
The pathways to net-zero



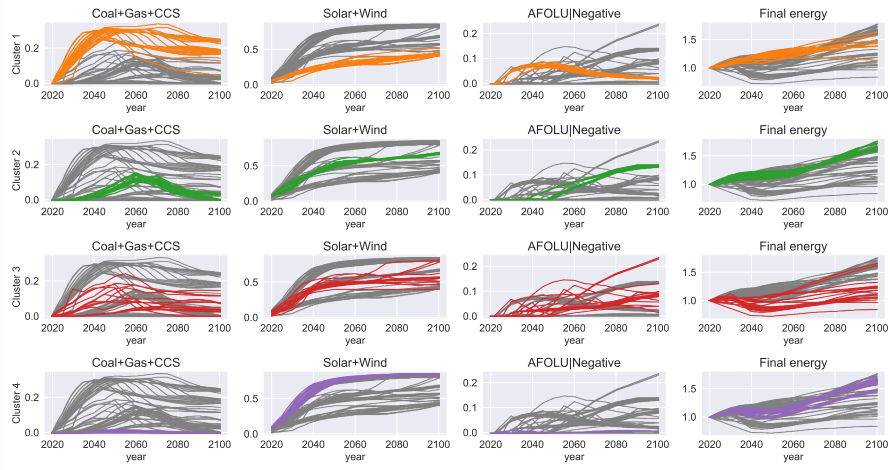
17 Marangoni et al. (in preparation)



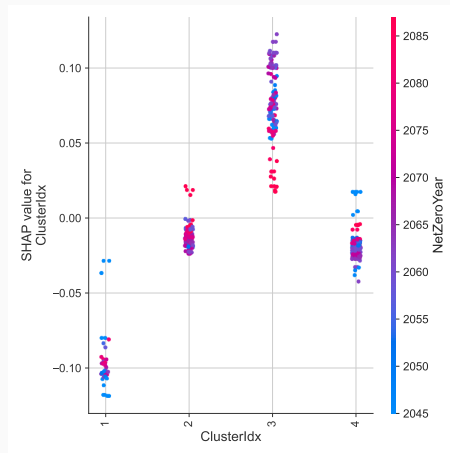
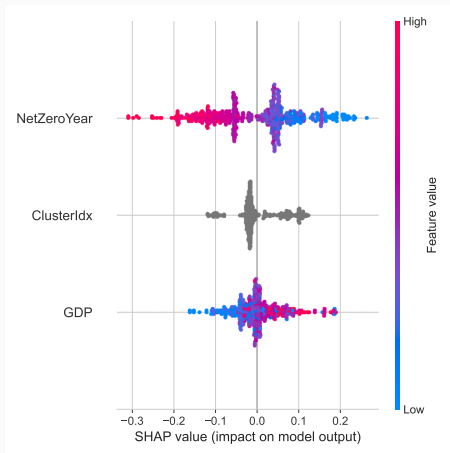
The pathways to net-zero



19 Marangoni et al. (in preparation)



20 Marangoni et al. (in preparation)



Marangoni et al. (in preparation)

Interpretable ML and beyond

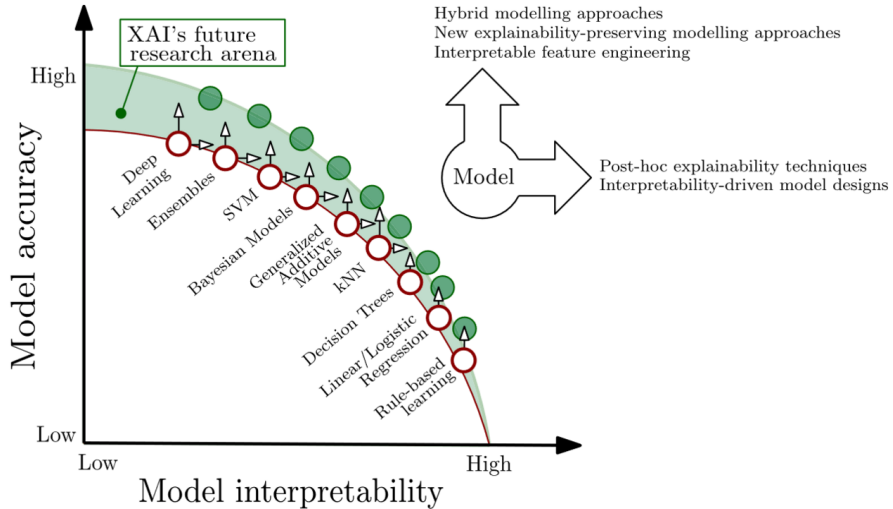
Guidelines for incorporating explainability in ML design

1. Consider contextual factors, impacts, and domain-specific needs;
2. Prefer interpretable techniques where possible;
3. If using a black-box model, ensure ethical, fair, and safe impacts are considered and mitigated with XAI tools;
4. Tailor interpretability to the human audience.

Post-hoc explainability limits



Trade-offs



Machine Learning

- [scikit-learn examples and documentation](#)
- [scikit-learn MOOC](#)
- [An Introduction to Statistical Learning Book](#)

Interpretable Machine Learning

- [Interpretable Machine Learning Book](#)

SHAP

- [SHAP documentation](#)

IAMs and ML

- Xiong et al. (2025). emIAM v1.0: An emulator for integrated assessment models using marginal abatement cost curves. *Geoscientific Model Development*, 18(5), 1575–1612.
<https://doi.org/10.5194/gmd-18-1575-2025>
- Shin et al. (2026). ML-IAM v1.0: Emulating Integrated Assessment Models With Machine Learning. *EGUsphere*, 1–24.
<https://doi.org/10.5194/egusphere-2025-5305>

Interpretable and explainable ML

- Lundberg et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), Article 1.
<https://doi.org/10.1038/s42256-019-0138-9>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), Article 5. <https://doi.org/10.1038/s42256-019-0048-x>

Econometrics and ML

- Mullainathan & Spiess (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87–106.
<https://doi.org/10.1257/jep.31.2.87>
- Athey & Imbens (2019). Machine Learning Methods That Economists Should Know About. Annual Review of Economics, 11(Volume 11, 2019), 685–725.
<https://doi.org/10.1146/annurev-economics-080217-053433>

ML and climate change

- Rolnick et al. (2022). Tackling Climate Change with Machine Learning. ACM Computing Surveys, 55(2), 42:1-42:96. <https://doi.org/10.1145/3485128>

Thank you!

Questions?

g.marangoni@tudelft.nl

This school received funding from the Swiss State Secretariat for Education, Research and Innovation (SEFRI) and the European Union through the Horizon Europe project PRISMA, “Net zero pathway research through integrated assessment model advancements” (grant no. 101081604), as well as from the Swiss National Science Foundation Eccellenza Grant for the project “Accuracy of long-range national energy projections” (grant no. 186834). I also gratefully acknowledge the funding of the European Research Council (ERC) under the European Union’s Horizon Europe research and innovation programme for the project RIPPLE (grant no. 101165221).

<https://mgiacomo.com/prisma-handson>

Appendix: Other post-hoc explanation methods

LIME: Local Interpretable Model-agnostic Explanations

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

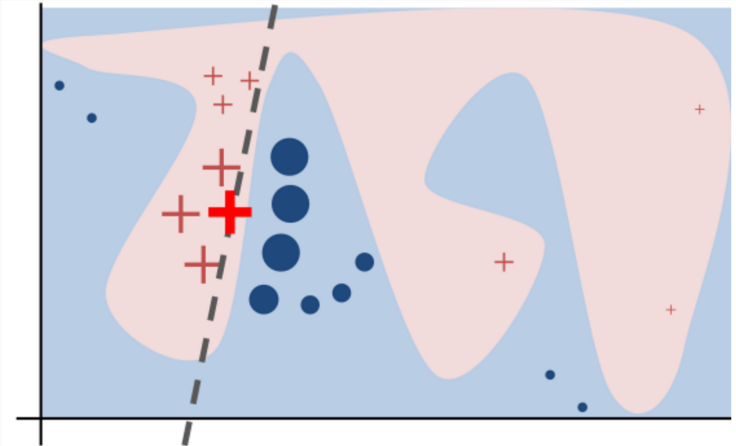
Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

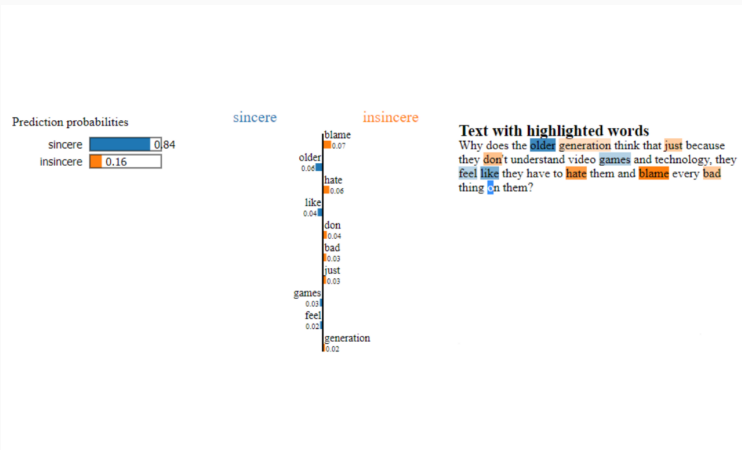
- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations.
- Explain the prediction by interpreting the local model.

LIME example

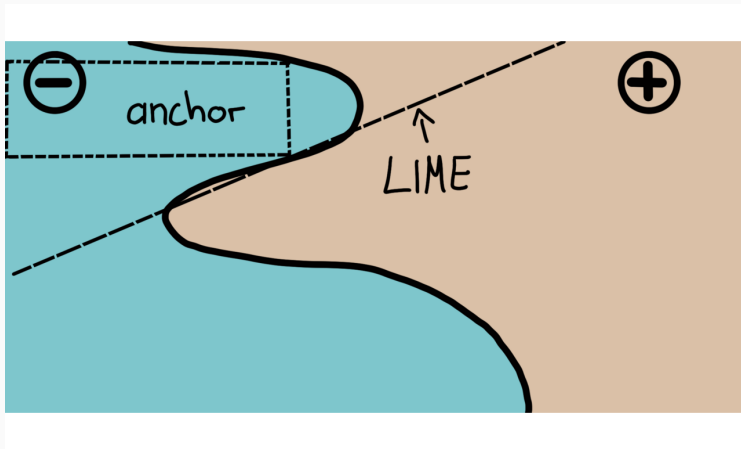


24 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). <https://doi.org/10.48550/arXiv.1602.04938>

LIME example



Anchors: High-Precision Model-Agnostic Explanations



26 <https://christophm.github.io/interpretable-ml-book/anchors.html>

Counterfactuals are alternative data points (x') that would change the model's prediction to a desired outcome (y').

Criteria for counterfactual selection

- Counterfactual instance prediction as close as possible to desired outcome
- Counterfactual instance as close as possible to the original instance
- Number of features changed from the original should be minimal
- Counterfactual instance as close as possible to observed data point

Dandl, S., Molnar, C., Binder, M., Bischl, B. (2020). https://doi.org/10.1007/978-3-030-58112-1_31

Counterfactual explanations

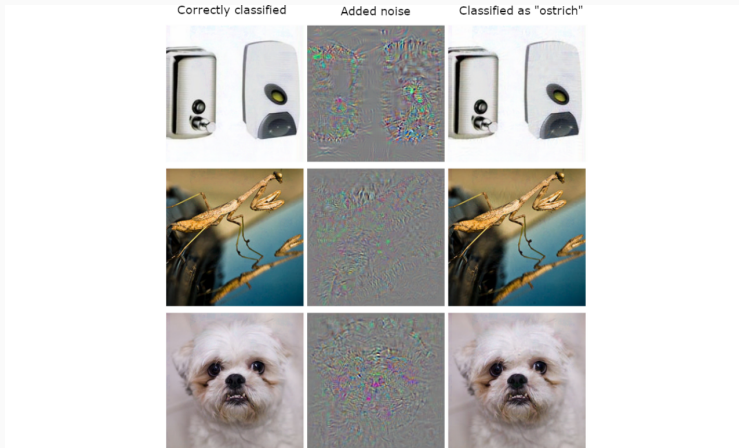
age	sex	job	housing	savings	amount	duration	purpose
58	f	unskilled	free	little	6143	48	car

The SVM predicts that the woman has a good credit risk with a probability of 24.2 %. The counterfactuals should answer how the input features need to be changed to get a predicted probability larger than 50 %?

age	sex	job	amount	duration	o_2	o_3	o_4	$\hat{f}(x')$
		skilled		-20	0.108	2	0.036	0.501
		skilled		-24	0.114	2	0.029	0.525
		skilled		-22	0.111	2	0.033	0.513
-6		skilled		-24	0.126	3	0.018	0.505
-3		skilled		-24	0.120	3	0.024	0.515
-1		skilled		-24	0.116	3	0.027	0.522
-3	m			-24	0.195	3	0.012	0.501

27 <https://christophm.github.io/interpretable-ml-book/counterfactual.html>

Adversarial examples



28 <https://christophm.github.io/interpretable-ml-book/adversarial.html>

Permutation feature importance

- Inputs: fitted predictive model m , tabular dataset (training or validation) D .
- Compute the reference score s of the model m on data D (for instance the accuracy for a classifier or the R^2 for a regressor).
- For each feature j (column of D):
 - For each repetition k in $1, \dots, K$:
 - Randomly shuffle column j of dataset D to generate a corrupted version of the data named $\tilde{D}_{k,j}$.
 - Compute the score $s_{k,j}$ of model m on corrupted data $\tilde{D}_{k,j}$.
 - Compute importance i_j for feature f_j defined as:

$$i_j = s - \frac{1}{K} \sum_{k=1}^K s_{k,j}$$

29 https://scikit-learn.org/stable/modules/permutation_importance.html