An Overview & Examples

Interpretable Machine Learning

2022-10-23

Outline

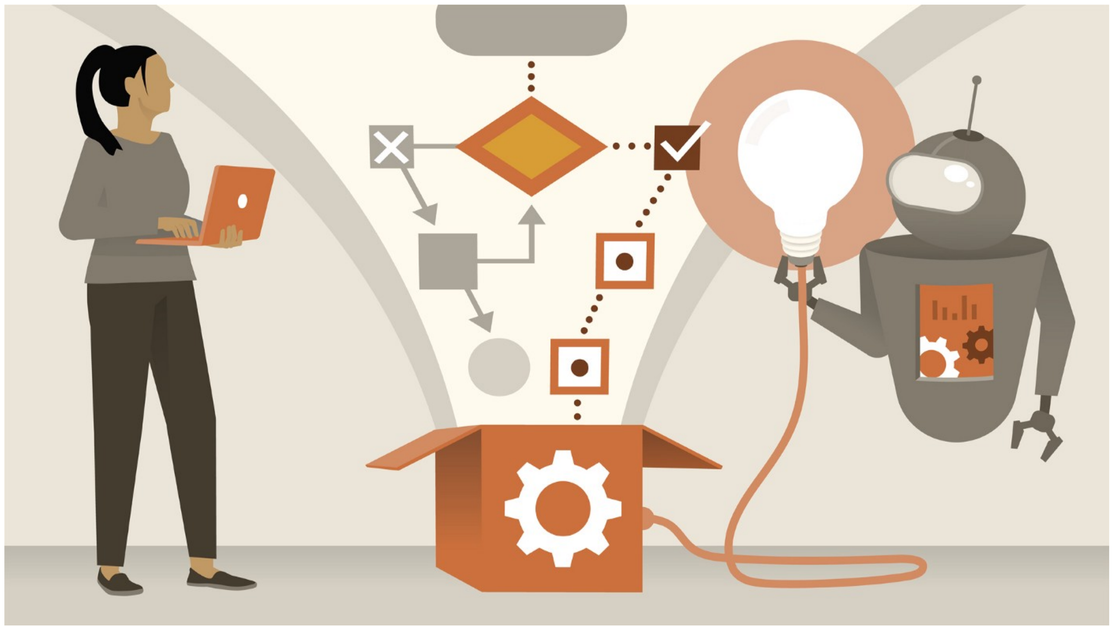
- Interpretable Machine Learning
- Taxonomy of Interpretable Machine Learning
- Global Model-Agnostic Method - **Permutation Feature Importance**
- Local Model-Agnostic Method - **LIME & SHAP**

Interpretable Machine Learning

What is Interpretable Machine Learning

- “(Models) are interpretable if their operations can be **understood by a human**, either **through introspection or through a produced explanation**.”

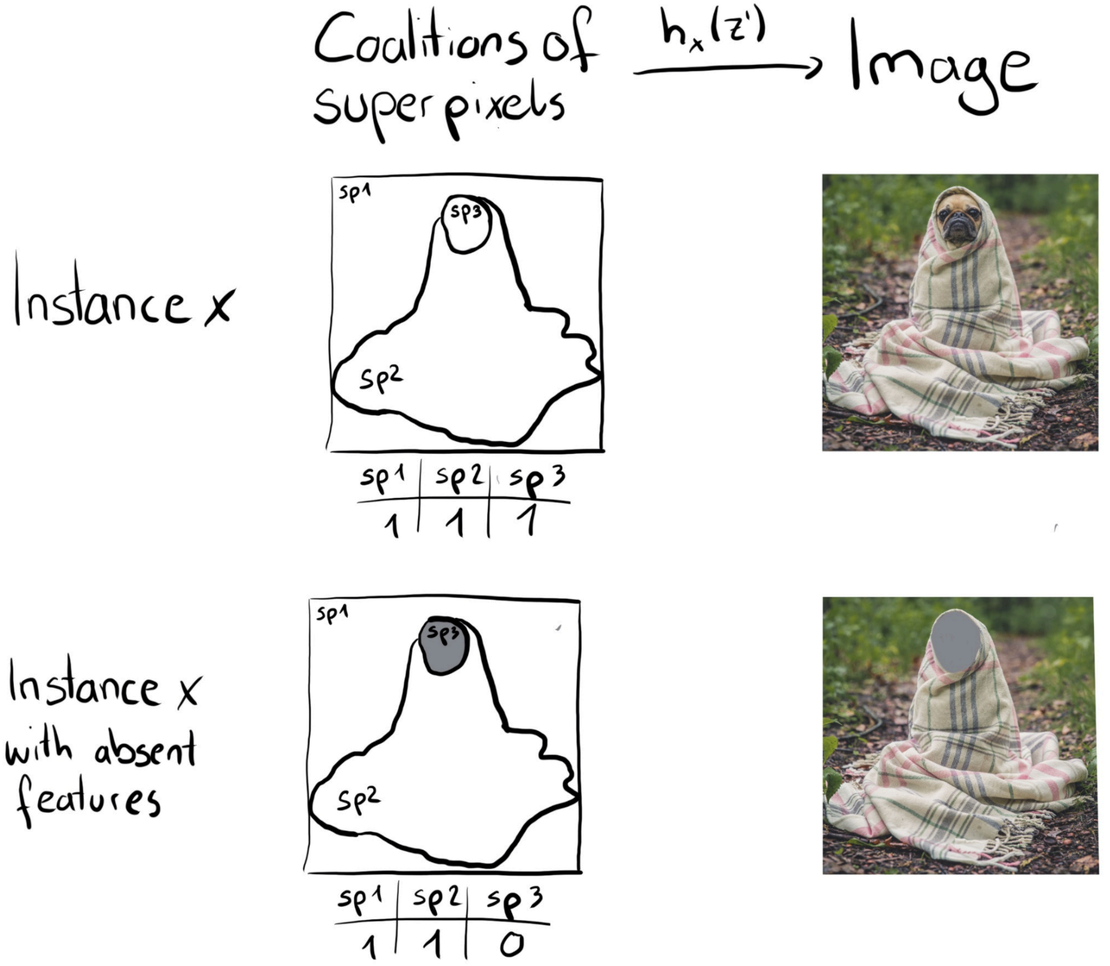
Biran, O., & Cotton, C. (2017, August). Explanation and justification in machine learning: A survey.



What is Interpretable Machine Learning

- Methods and Models that make the **behavior and predictions** of Machine Learning Systems understandable to humans.

Molnar, C. (2020). Interpretable machine learning.



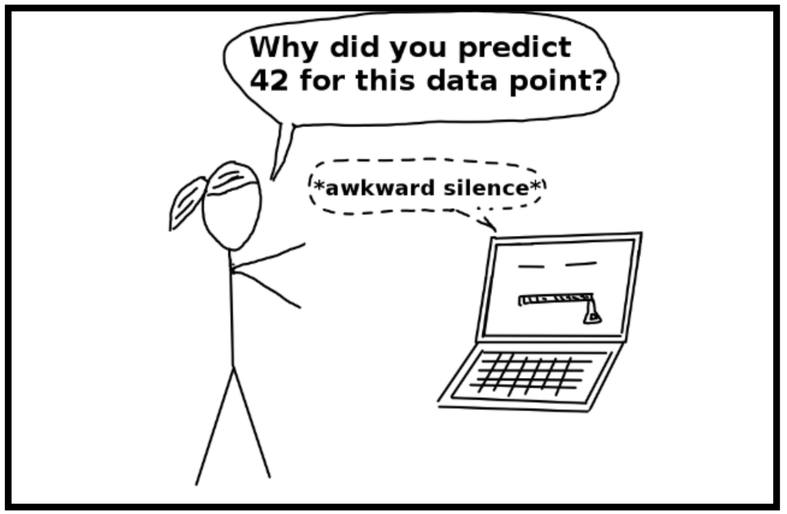
Interpretability vs Explanation

✓ The degree to which a human can consistently predict the model's result.

Interpretability

Model's Behavior

Explanation



Individual Prediction

Why Interpretable Machine Learning?

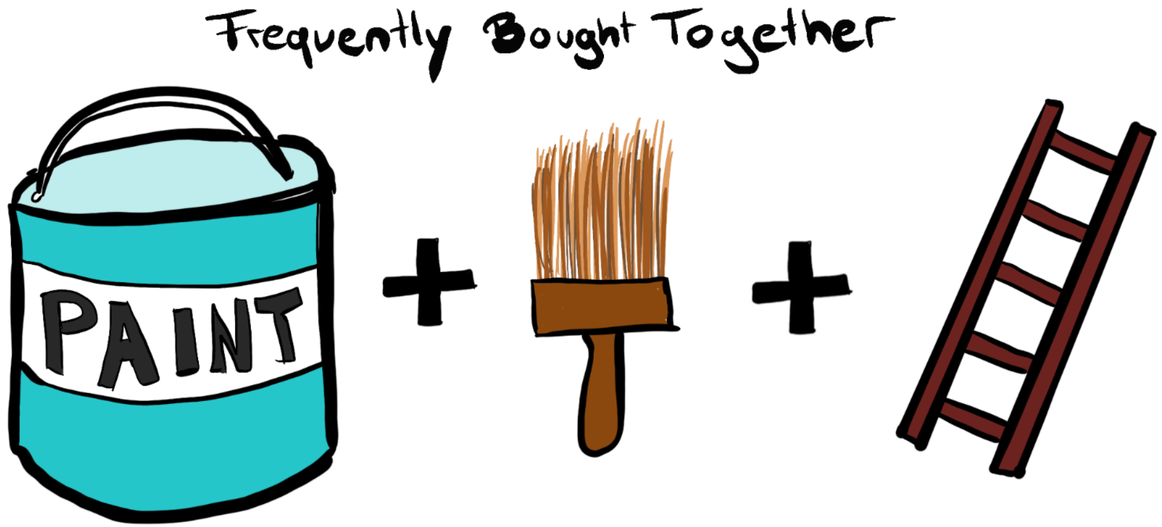
Why should I trust you?

- Why I feel so sick?
- Why did not the treatment work on patient?
- Why was my loan rejected?
- Why this product recommended to me?
- Why this student couldn’t pass this exam and how can we help him?

Why Interpretable Machine Learning?

Human curiosity, learning and find meaning the world

- The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior.

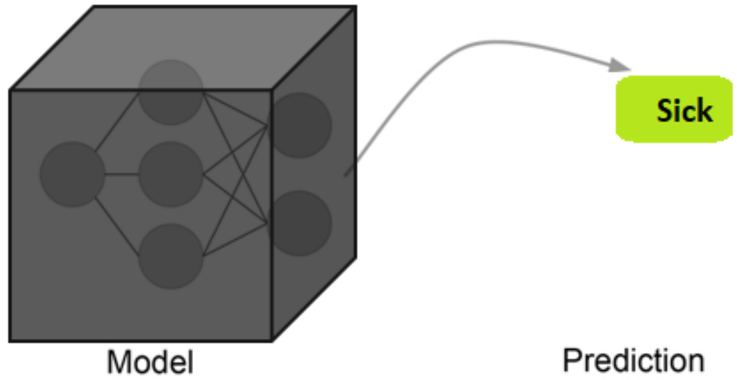


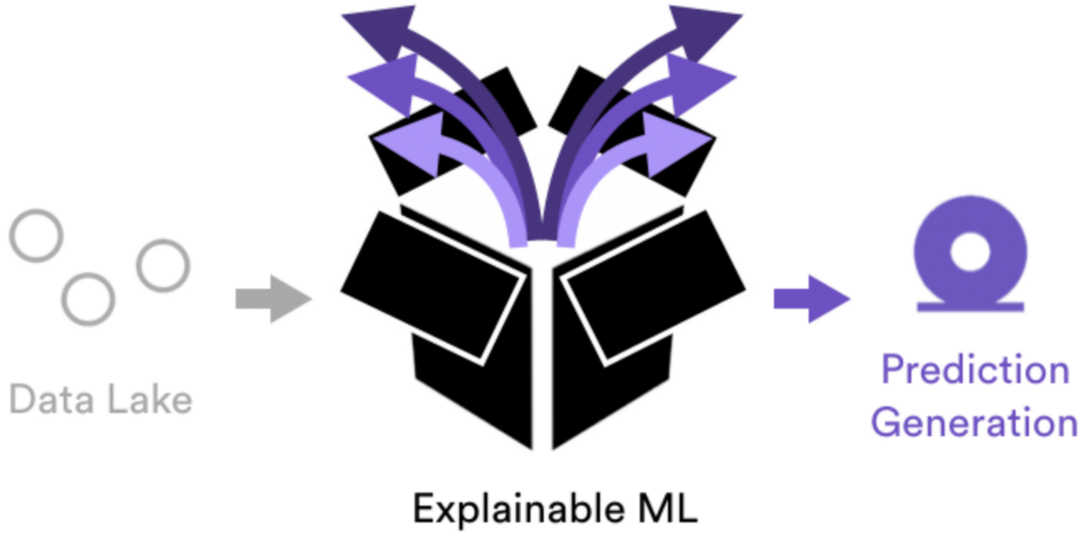
Why Interpretable Machine Learning?

The goal of science is to gain knowledge

- With Big Data, the model itself becomes the source of knowledge instead of the data.

- Interpretability makes it possible to extract this additional knowledge captured by the model.

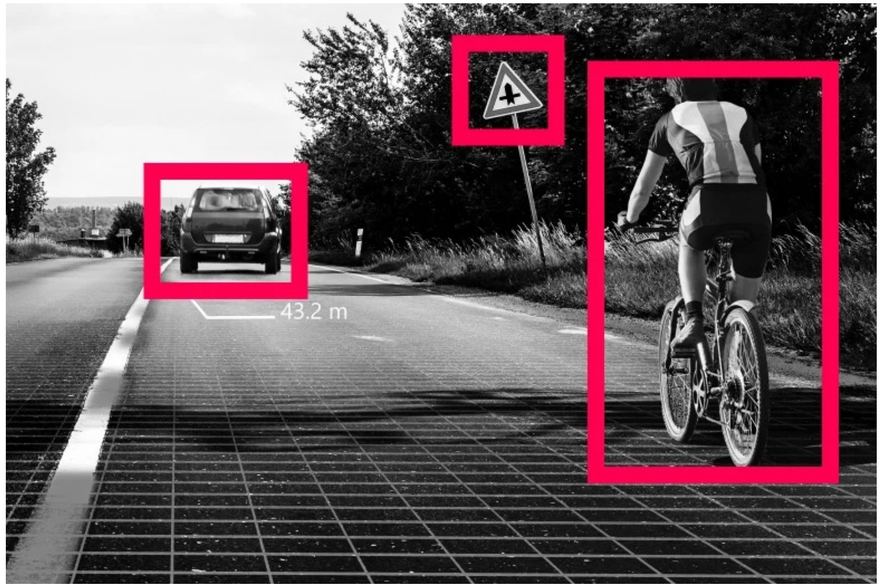


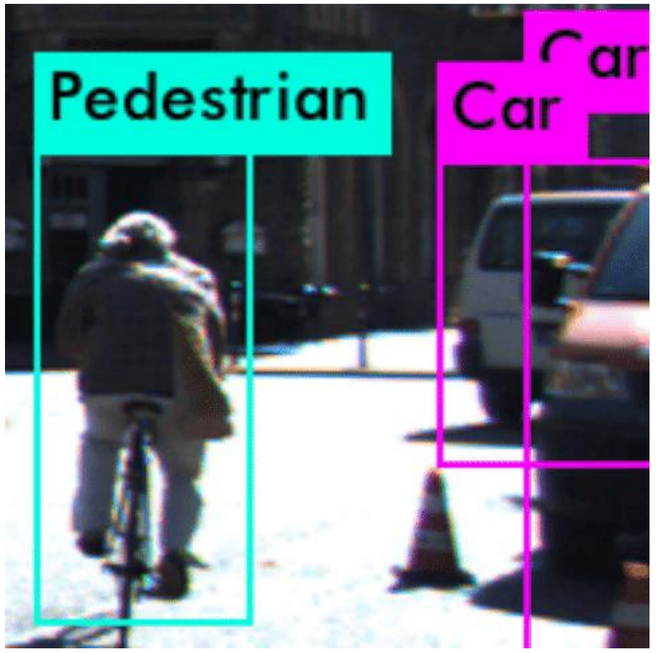


Why Interpretable Machine Learning?

The safety measure in real-world task

- Explanation might reveal the most importance learned feature is recognize, help to think about edge cases.





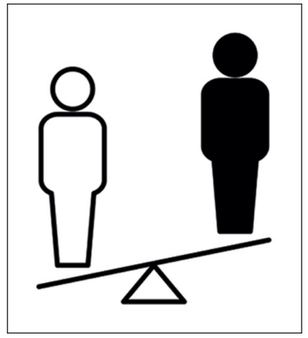
Why Interpretable Machine Learning?

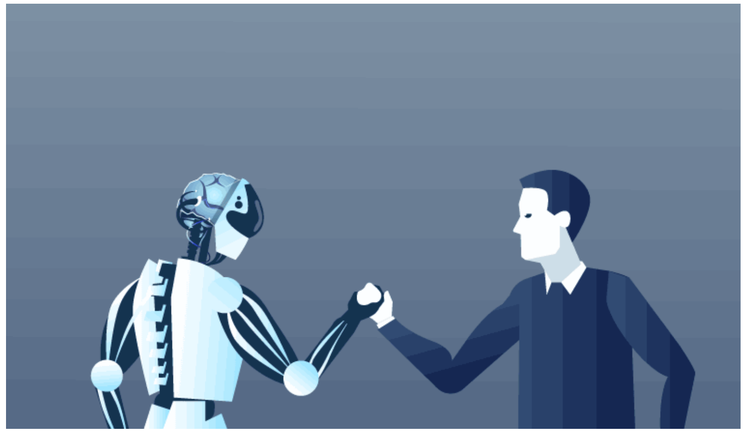
Detecting bias

Social acceptance

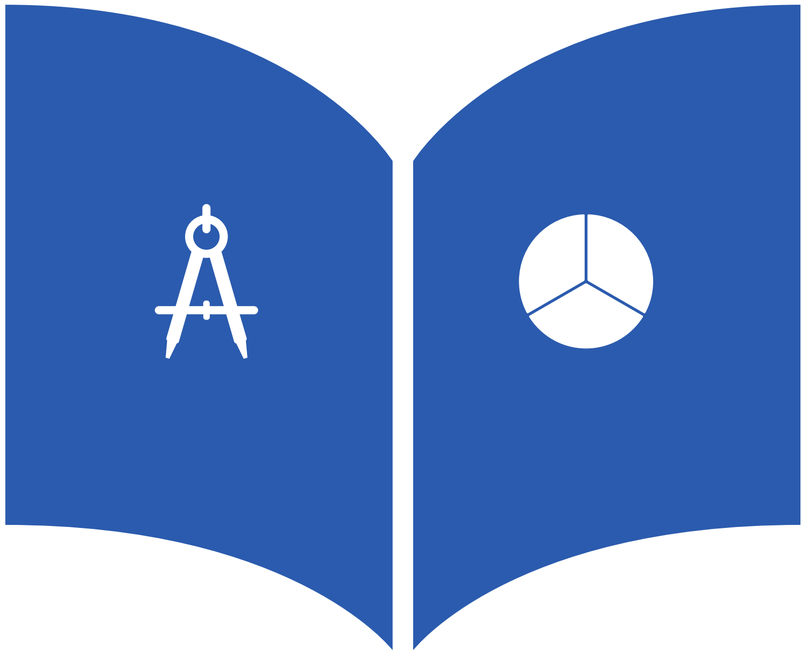
Manage social interactions

Debugged and audited





Taxonomy of Interpretability Methods

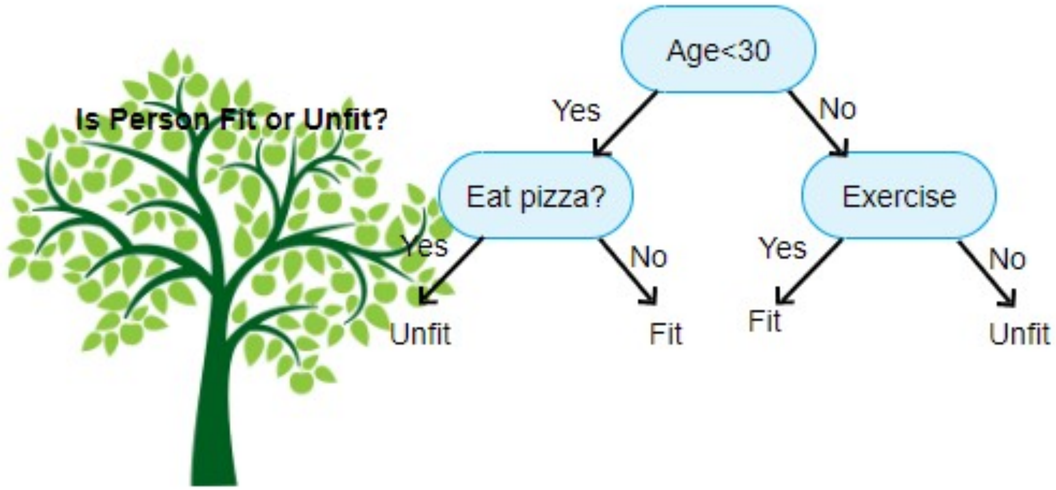


- Intrinsic or Post-hoc
- Result of Interpretation Method
- Model-Specific or Model-Agnostic
- Local or Global

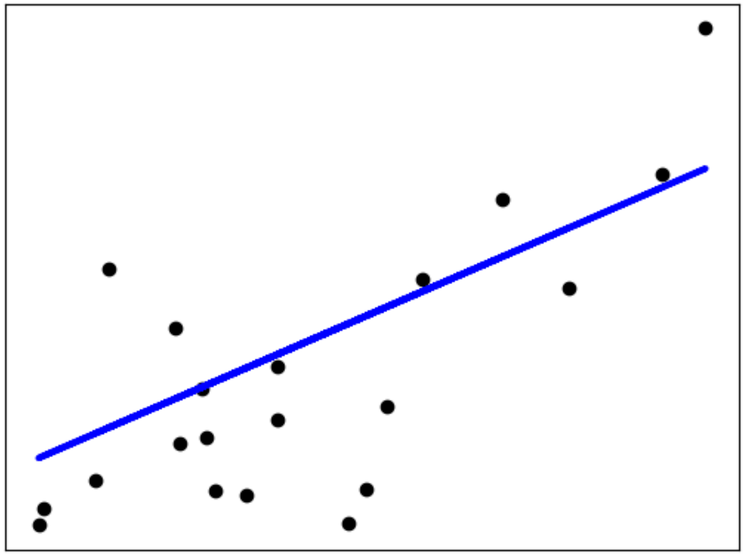
Taxonomy of Interpretability Methods

Intrinsic or Post-hoc

- **Intrinsic**: Interpretable due to their simple structure



Decision Tree

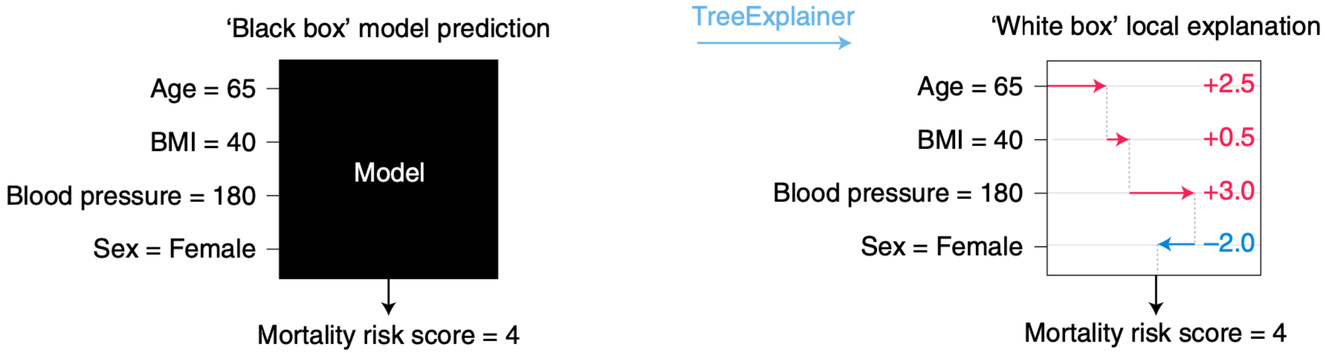


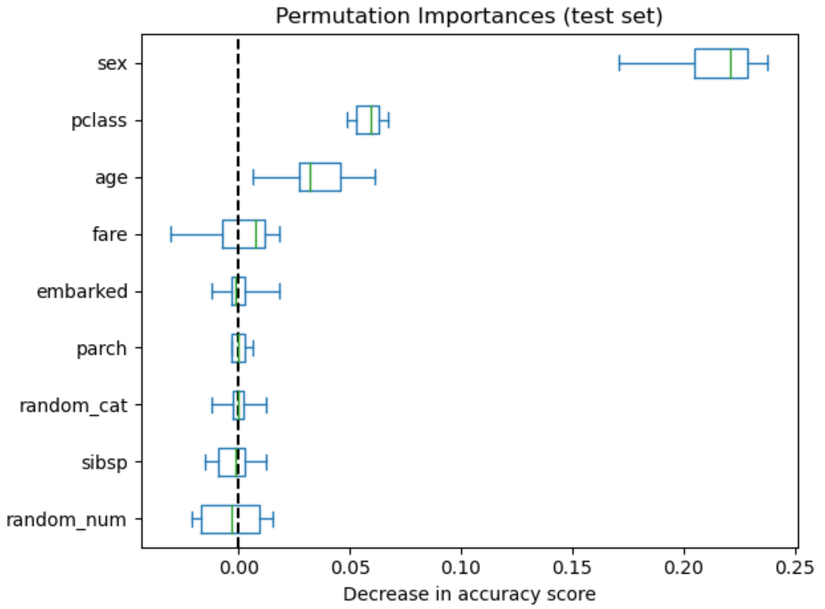
(Sparse) Linear Regression

Taxonomy of Interpretability Methods

Intrinsic or Post-hoc

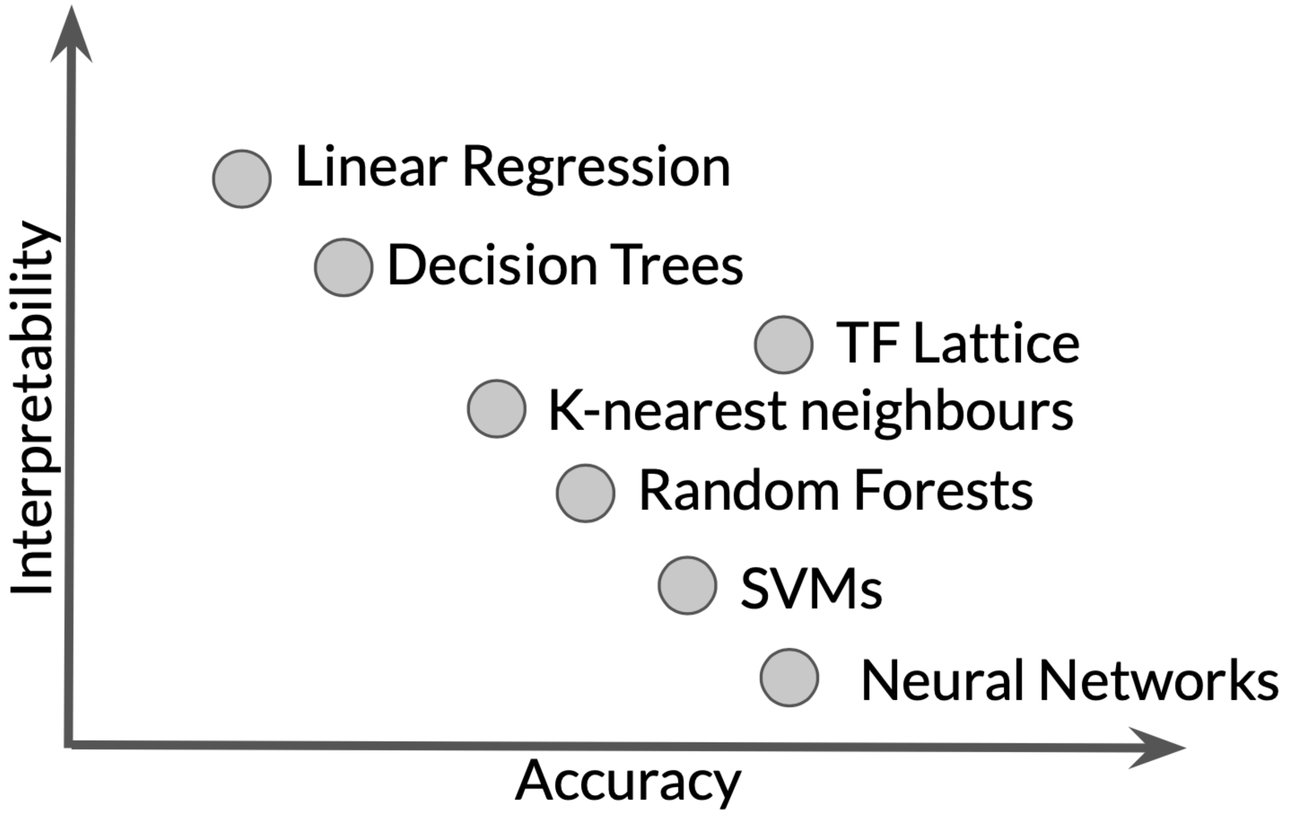
- **Post-hoc**: Application of interpretation after training
- Treat mode as black boxes
- Agnostic to mode architecture
- Extract relationship between input & prediction





Taxonomy of Interpretability Methods

Intrinsic or Post-hoc



Interpretability vs Accuracy Trade off

Taxonomy of Interpretability Methods

Result of Interpretation Method

Feature Summary Statistics

Feature Summary Visualization

Model internals *(learned weights):*

weights in linear models or learned tree structures

(features and threshold used for split)

Model Internals

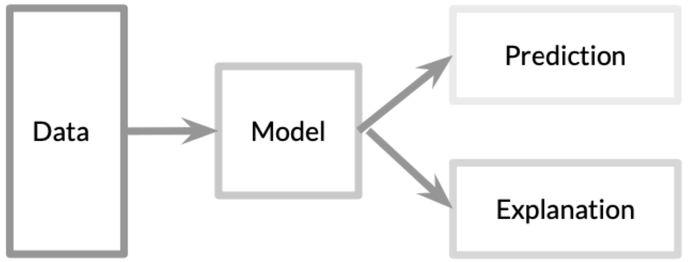
Data point

Taxonomy of Interpretability Methods

Model-Specific or Model-Agnostic

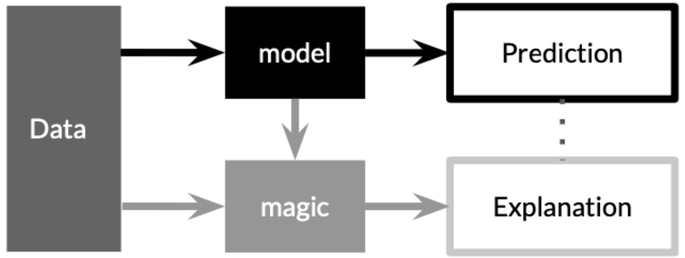
Model Specific

- These tools are limited to specific model classes
- Example: Interpretation of regression weights in linear models
- Intrinsically interpretable model techniques are model specific
- Tools designed for particular model architectures



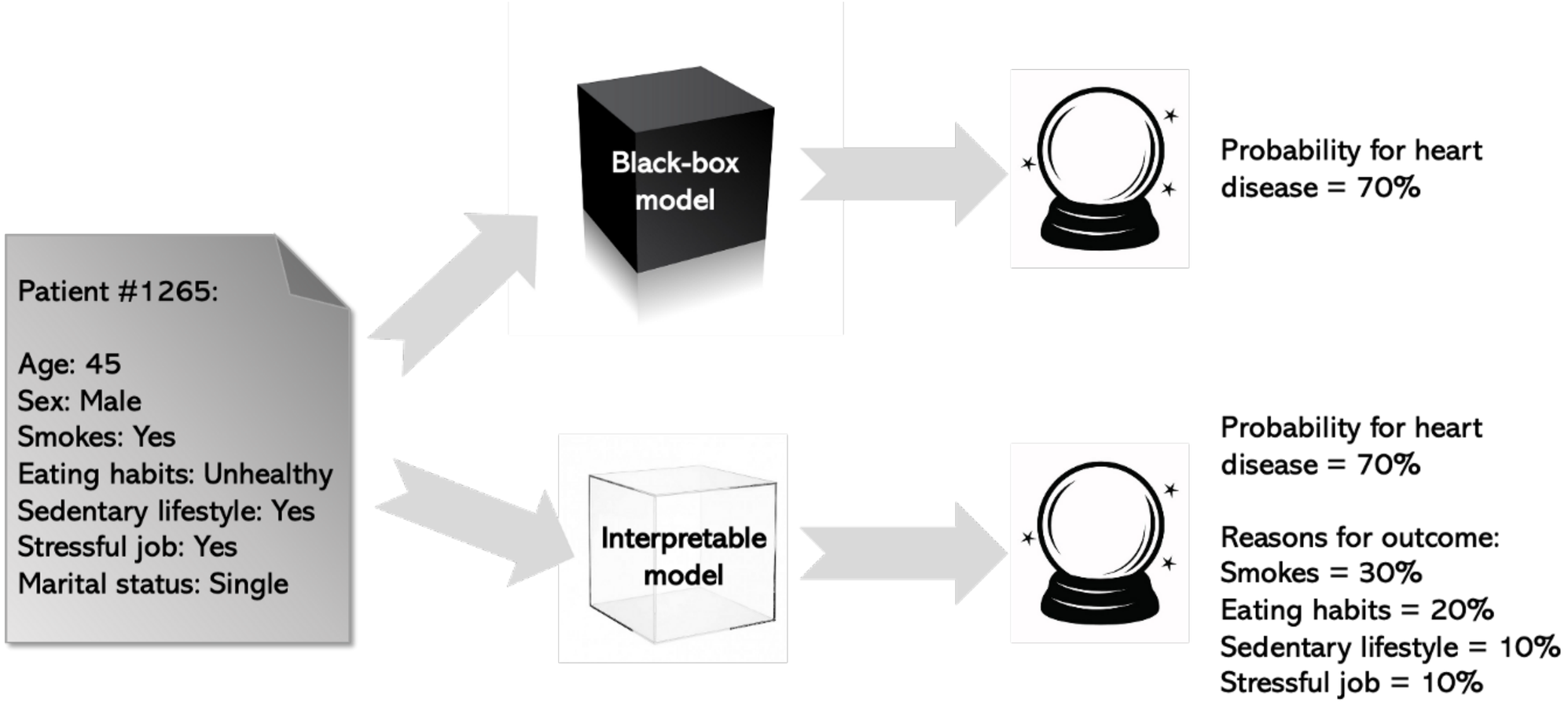
Model Agnostic

- Applied to any model after it is trained
- Do not have access to the internals of the model
- Work by analyzing feature input and output pairs



Taxonomy of Interpretability Methods

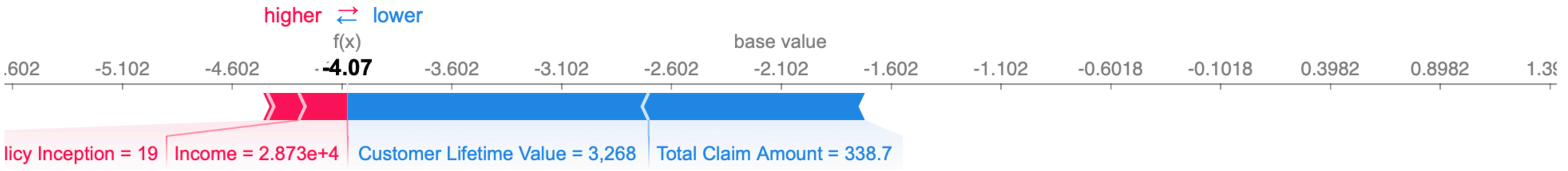
Model-Specific or Model-Agnostic



Taxonomy of Interpretability Methods

Local or Global

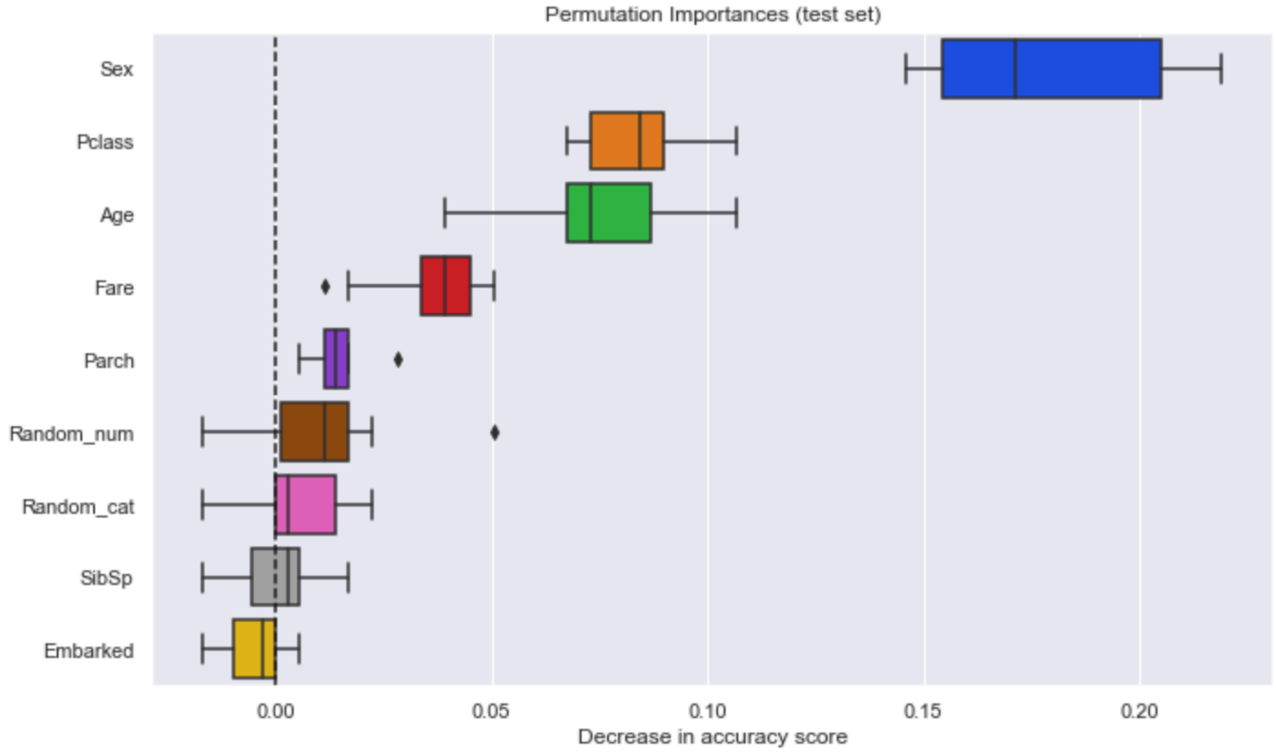
- **Local:** interpretation method explains *individual prediction*



Taxonomy of Interpretability Methods

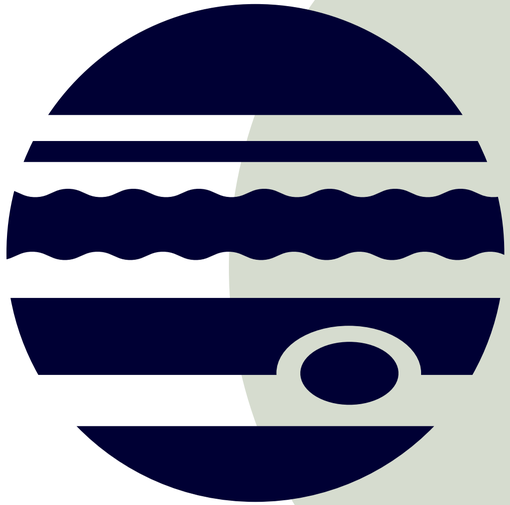
Local or Global

- **Global:** interpretation method explains *entire model behaviour*



- Global methods describe how features affect the prediction **on average.**

Global Model-Agnostic Methods



Partial Dependence Plot (PDP)

Feature Interaction (H-statistic)

Functional Decomposition

Permutation Feature Importance

Global Surrogate Model

Prototypes and Criticism

Global Model-Agnostic Methods

Permutation Feature importance

Measures the increase in the prediction error of the model after permuting the feature's values:

- **Important** features:

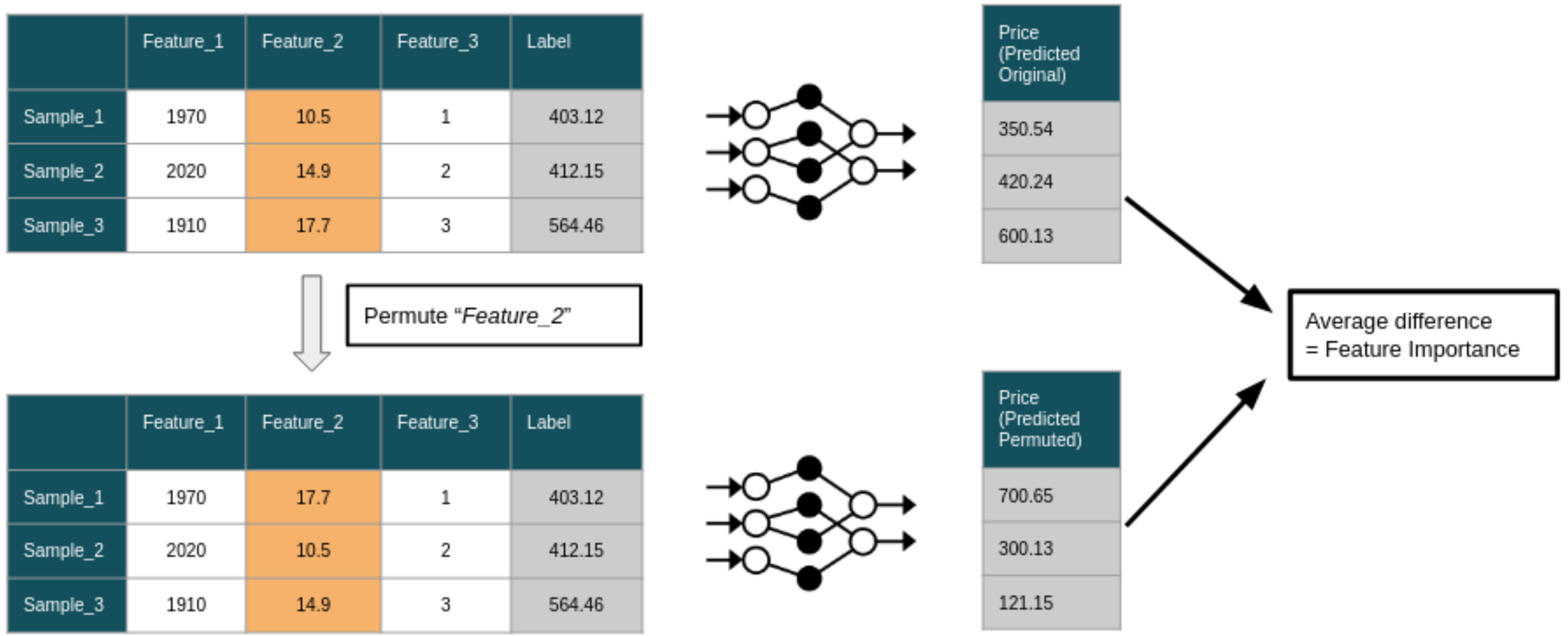
  Model error increases

  when shuffling it

- **Unimportant** features:

  Model error unchanged

  when shuffling it



Global Model-Agnostic Methods

Permutation Feature importance

f: trained model **X(m, n)**: feature matrix

y: target vector **L(y, f)**: error measure

For each feature j ∈ {1, ..., n}:

- Generate feature X_{perm} by permuting feature matrix X
- Estimation error $e_{perm} = L(Y, f(X_{perm}))$
- Calculate permutation feature importance $FI_j = e_{perm} / e_{org}$ or $FI_j = e_{perm} - e_{org}$

Sort by descending FI

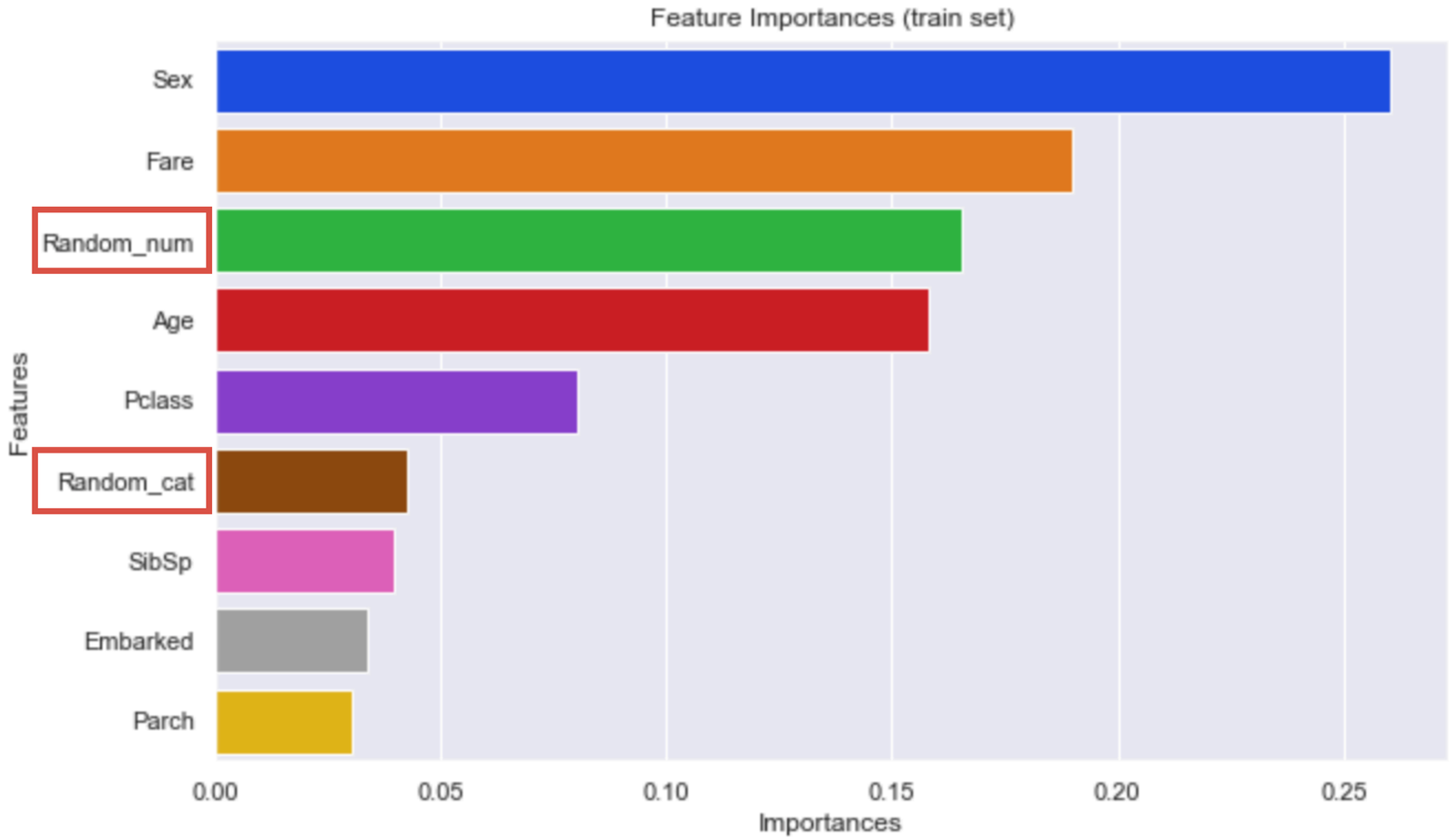
Global Model-Agnostic Methods

Permutation Feature importance vs **Random Forest Feature Importance**

- **Titanic** dataset
- Including two random variables that are not correlated in any way with the target variable.

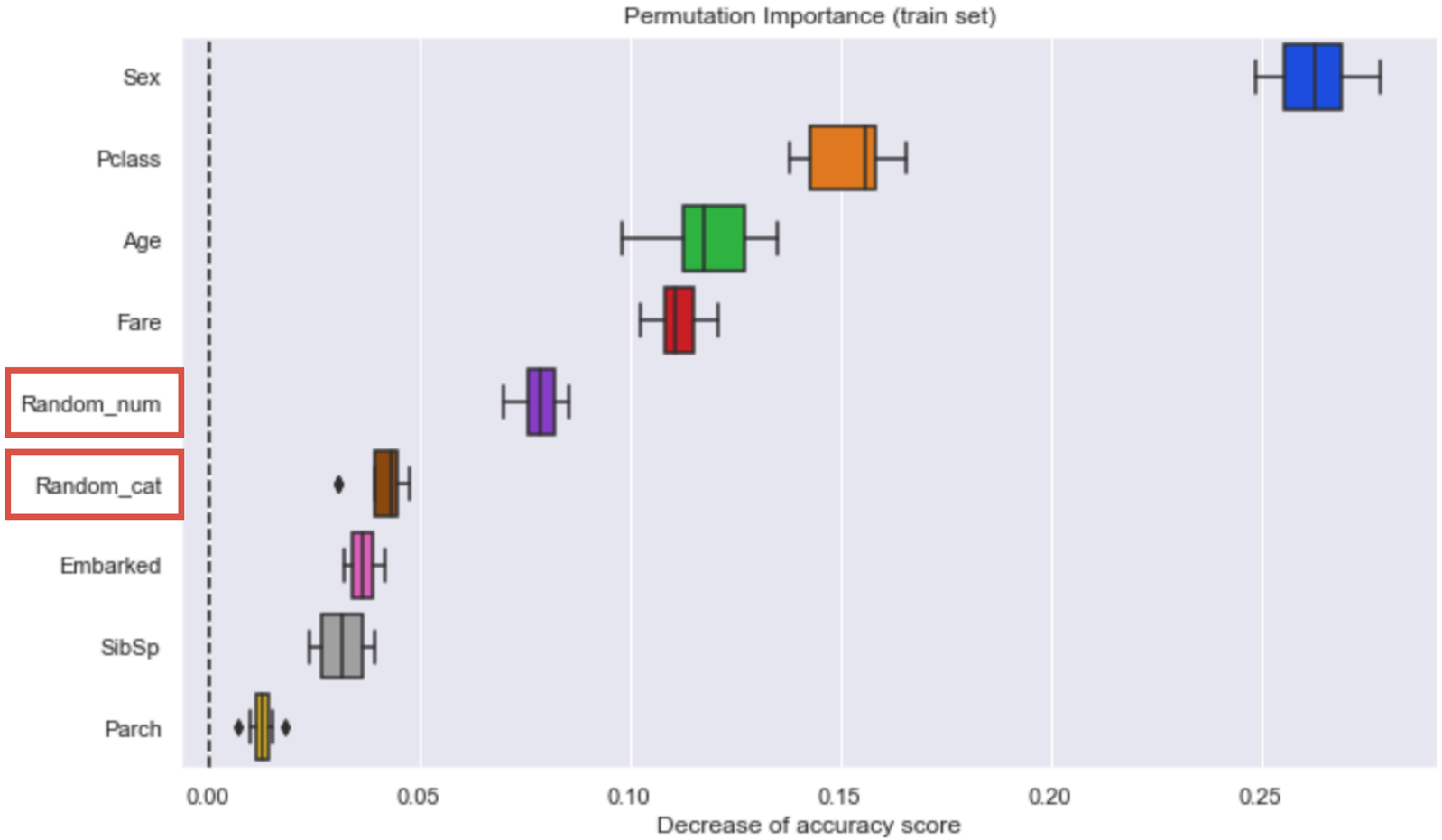
Random_num: numerical variable (many unique values).

Random_cat: categorical variable (3 possible values).



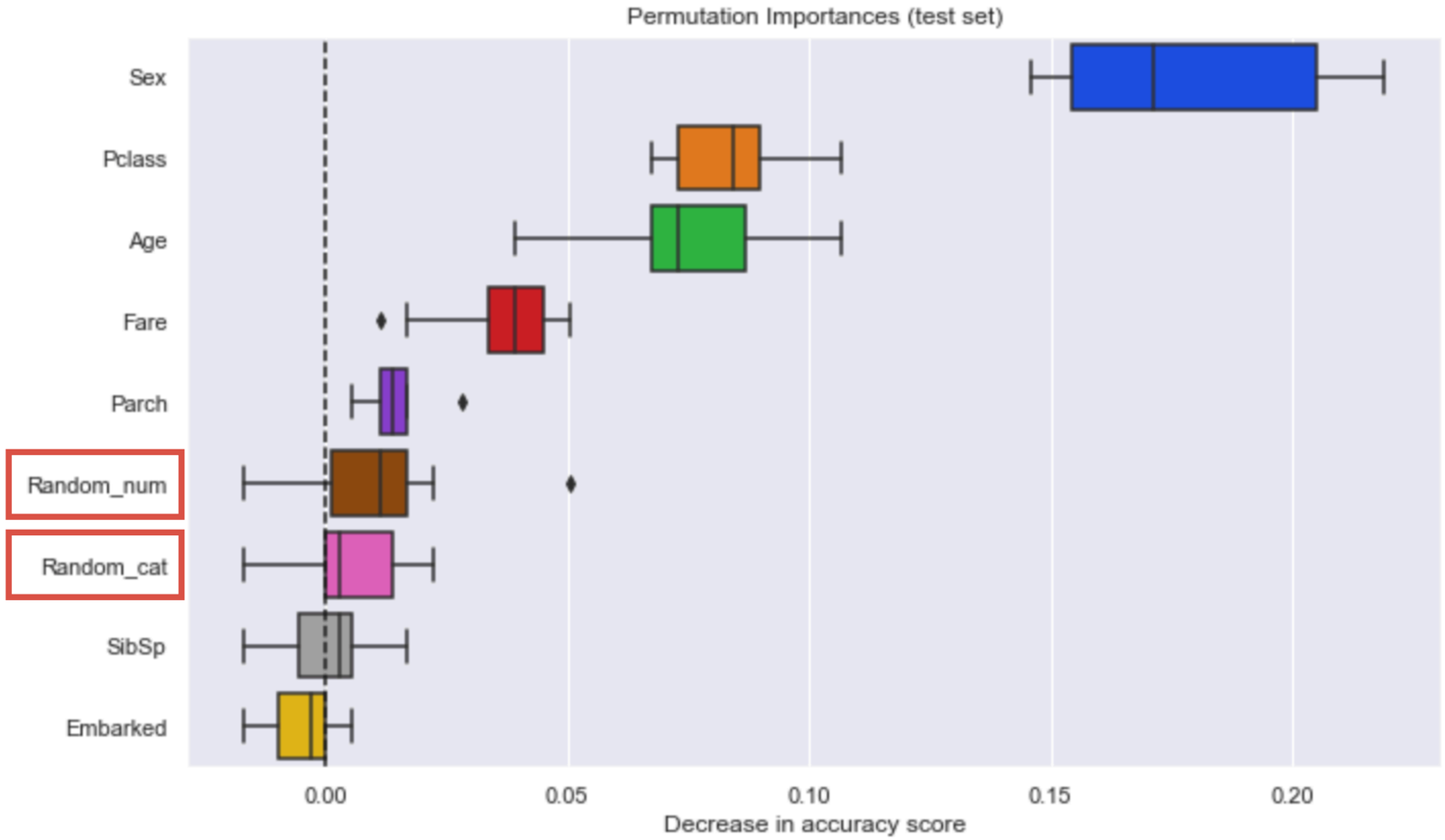
Global Model-Agnostic Methods

Permutation Feature importance vs Random Forest Feature Importance



Global Model-Agnostic Methods

Permutation Feature importance vs Random Forest Feature Importance



Global Model-Agnostic Methods

Permutation Feature importance

Advantages

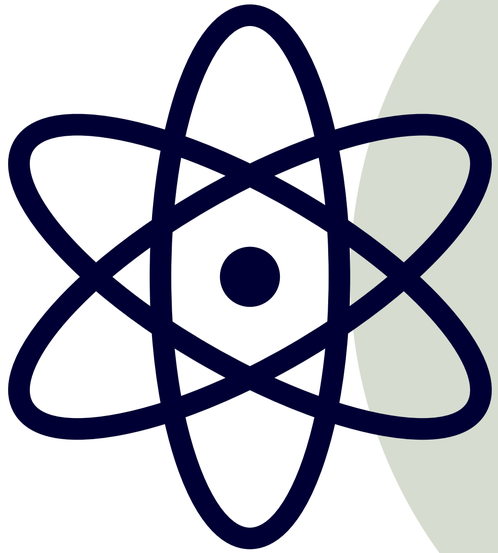
- Nice interpretation
- Provide highly compressed, global insight into the model behavior
- Does not required retraining model

.

Disadvantages

- It is unclear if testing or training data should be used for visualization.
- Can be biased by unrealistic data instances
- Adding a correlated feature can decrease the importance of the associated feature
- Need access to the true outcome

Local Model-Agnostic Methods



Individual Conditional Expectation (ICE)

Local Surrogate (LIME)

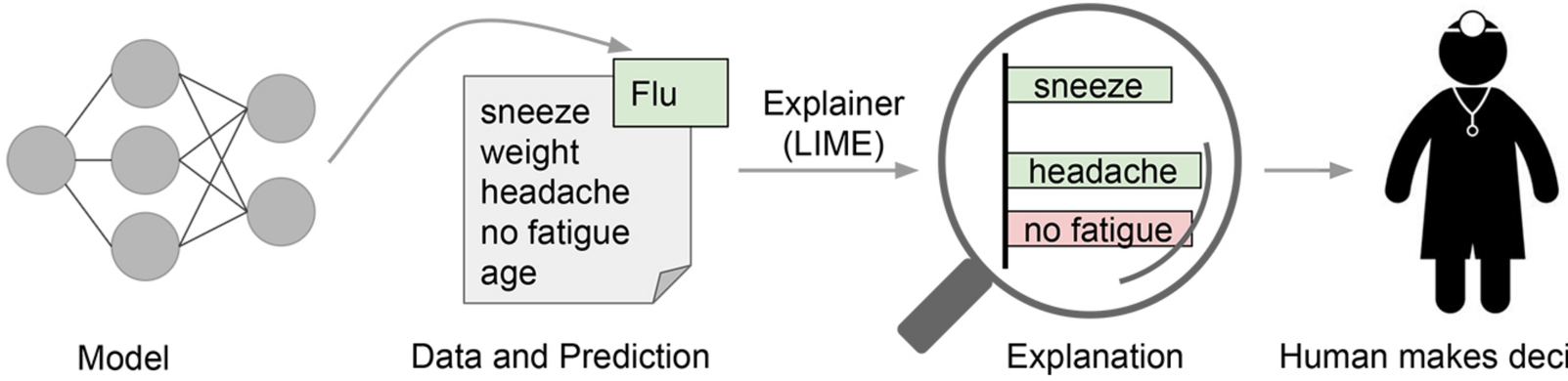
Counterfactual Explanations

Scoped Rules (Anchors)

Shapley Values / SHAP (SHapley Additive exPlanations)

Local Model-Agnostic Methods

LIME - Local Surrogate



- Surrogate models are trained to **approximate** the predictions of the underlying black box model. Instead of training a global surrogate model.
- LIME focuses on training local surrogate models to **explain individual predictions**.
- Using data points close to the individual prediction, LIME trains an interpretable model to **approximate** the predictions of the real model.
- The new interpretable model is then used to interpret the real result.

Local Model-Agnostic Methods

LIME | Calculation

Model complexity

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Family of interpretable models

Complex models

Simple interpretable model

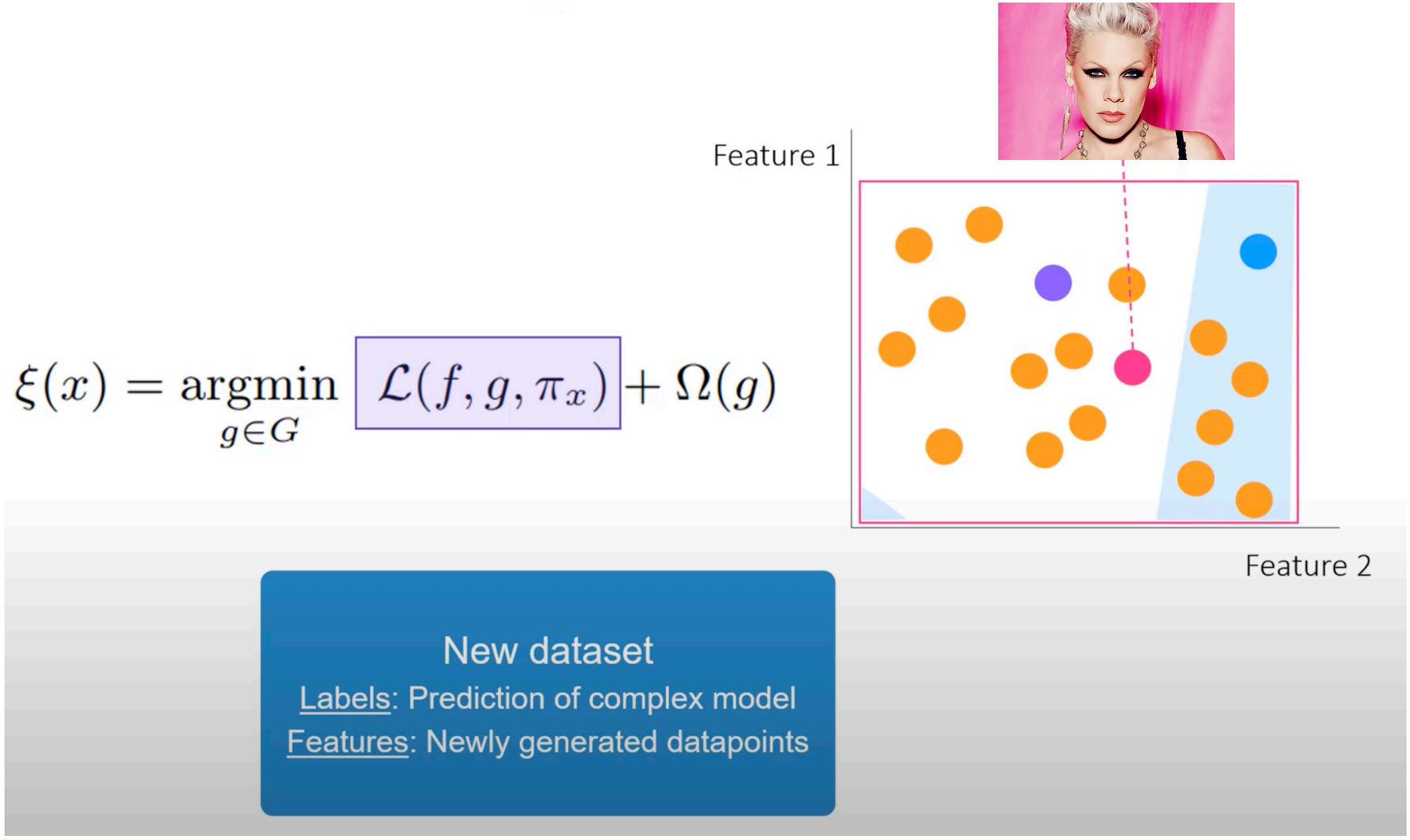
Proximity

Local Model-Agnostic Methods

LIME - The recipe

- Select instance
- Perturb dataset and get new predictions.
- Weight new samples by their proximity to instance.

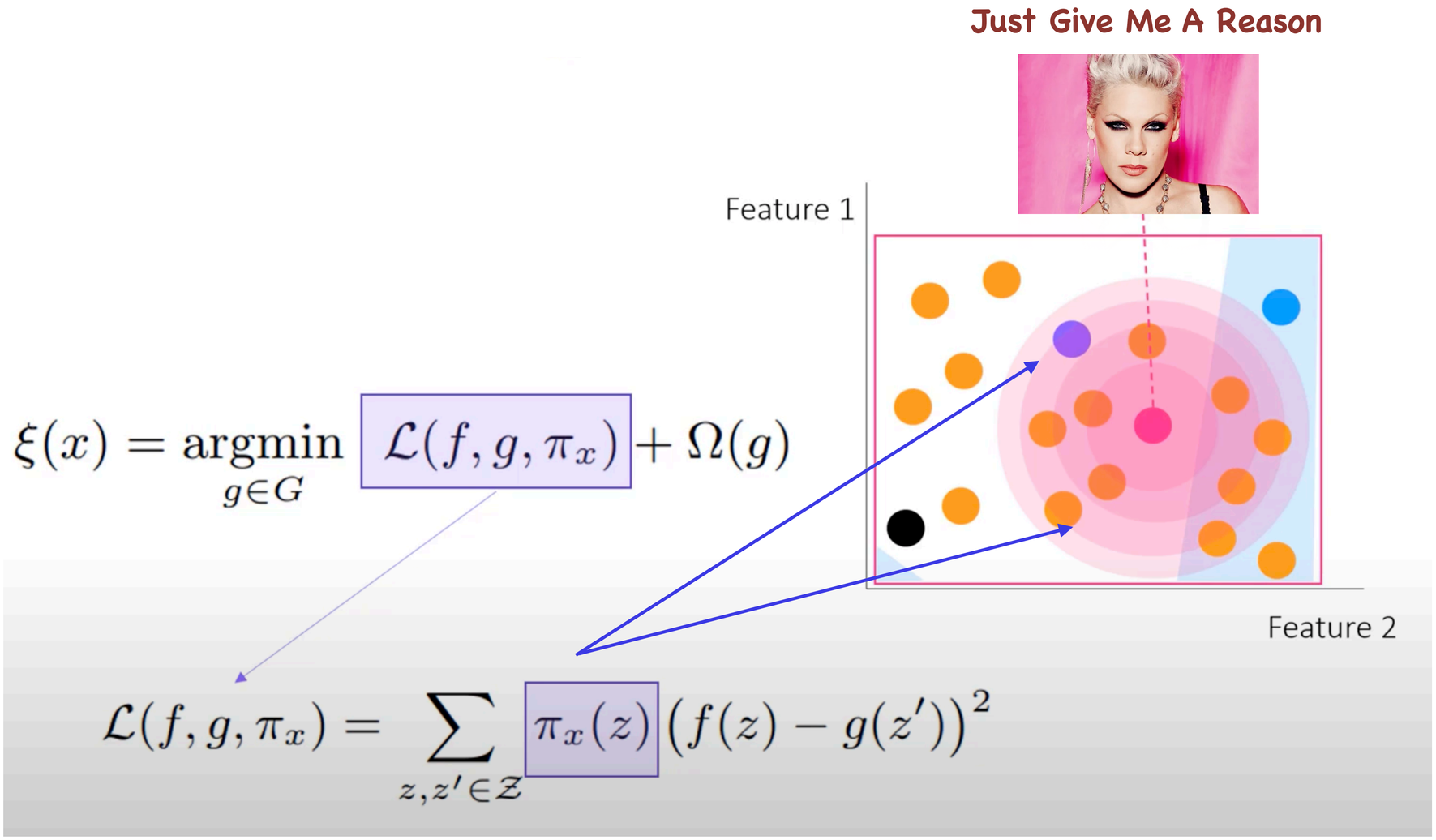
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



Local Model-Agnostic Methods

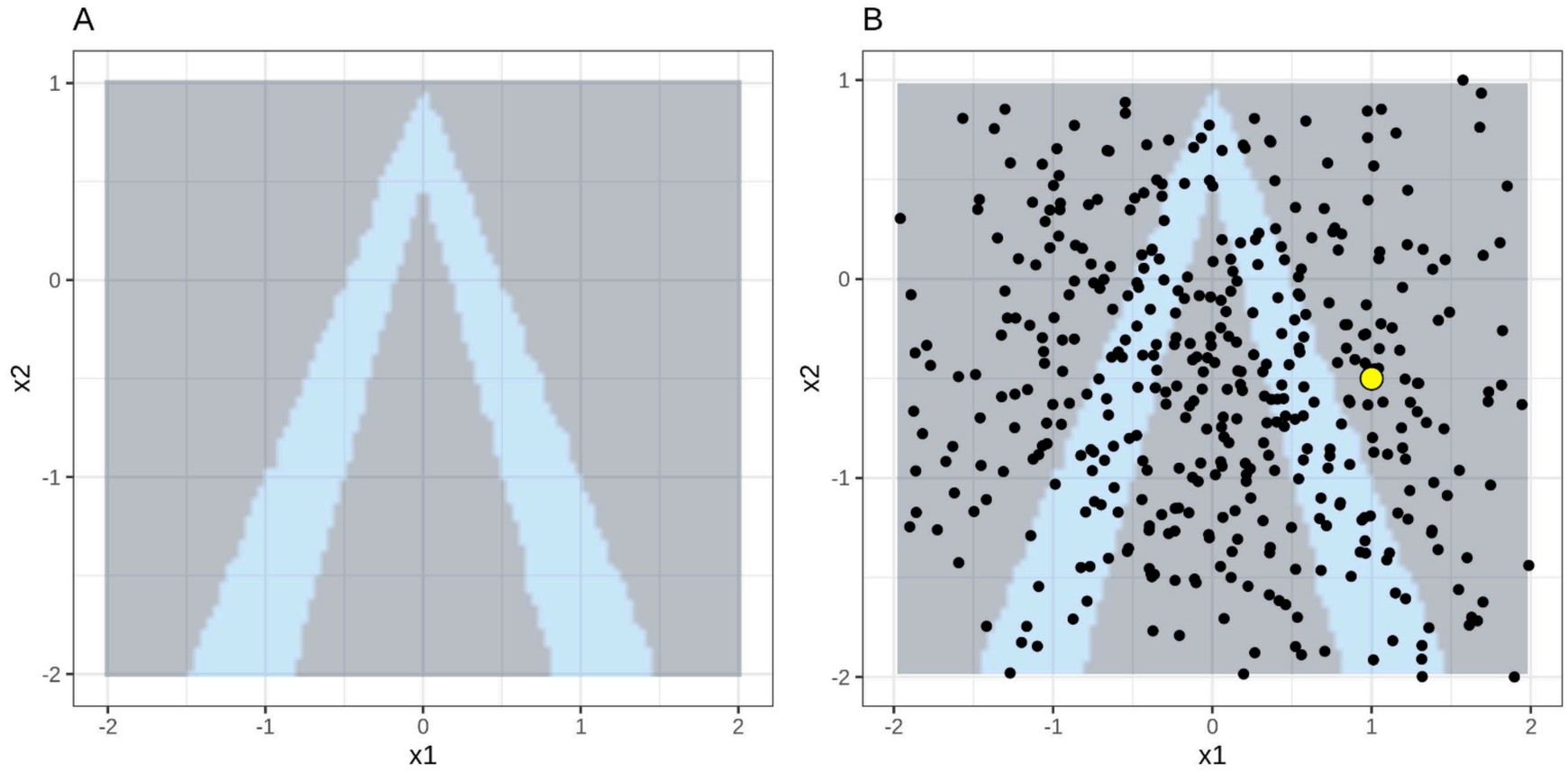
LIME - The recipe

- Trained a weighted, interpretable model on new dataset
- Explain prediction by interpreting local model.



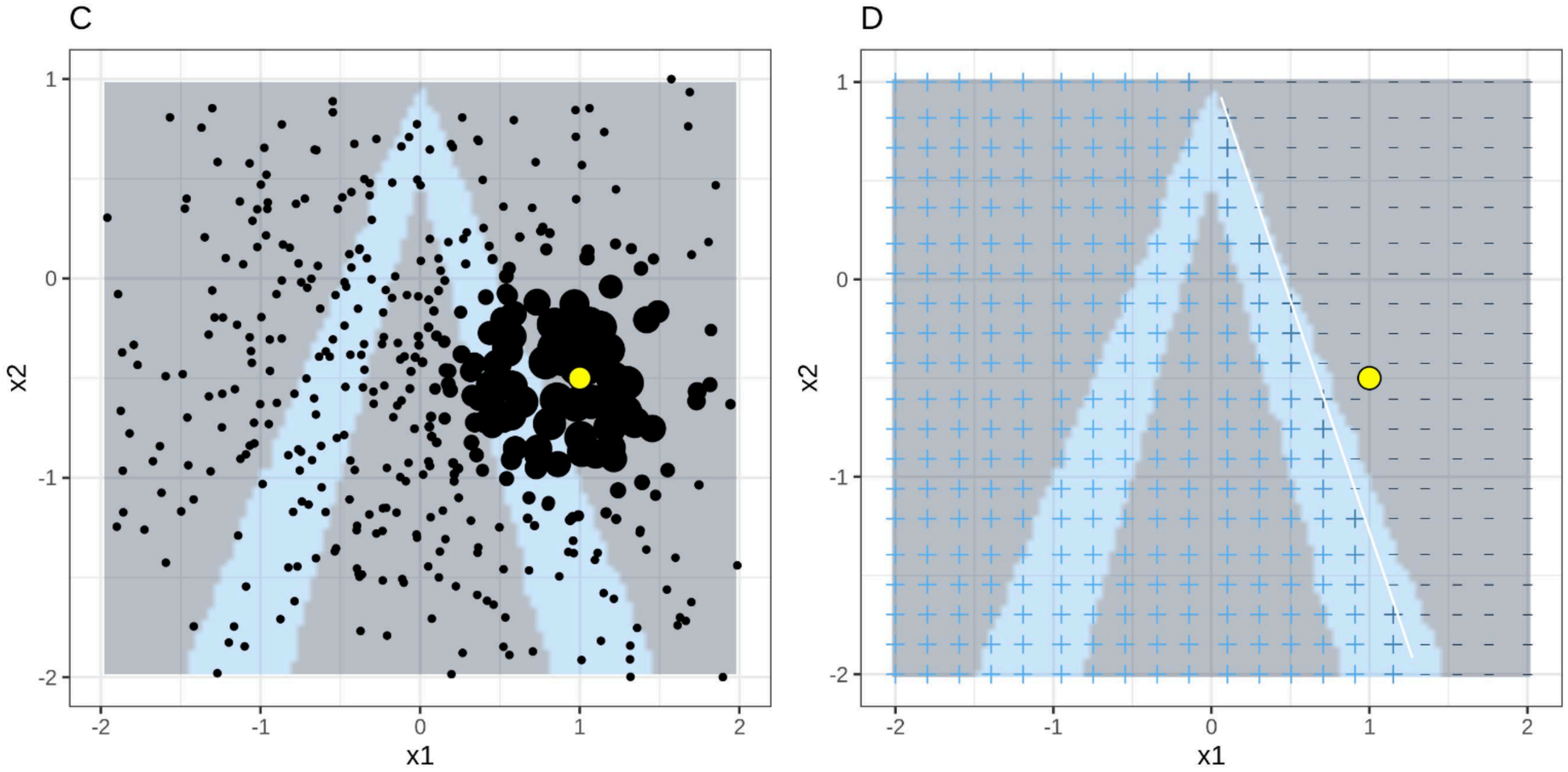
Local Model-Agnostic Methods

LIME - Tabular Data



Local Model-Agnostic Methods

LIME - Tabular Data



Local Model-Agnostic Methods

LIME - Tabular Data | Classification

	Customer Lifetime Value	Income	Monthly Premium Auto	Months Since Policy Inception	Total Claim Amount	Response
9129	23405.987980	71941	73	89	198.234764	0
9130	3096.511217	21604	79	28	379.200000	1
9131	8163.890428	0	85	37	790.784983	0
9132	7524.442436	21941	96	3	691.200000	0
9133	2611.836866	0	77	90	369.600000	0

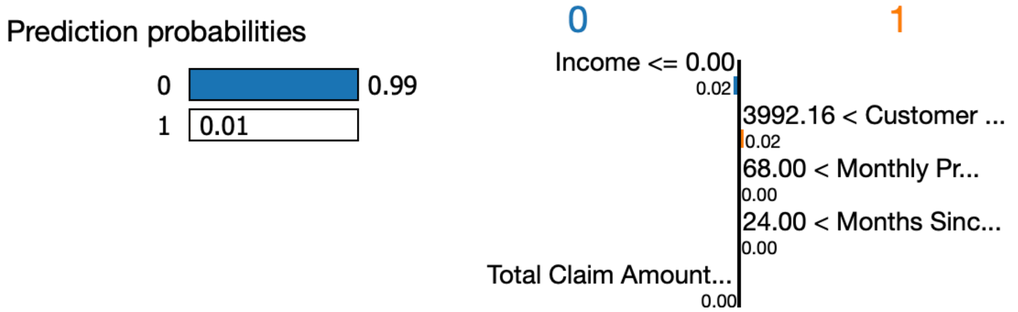
- **Data:** Marketing-Customer-Value-Analysis
- **Label:** Customer response (1) and not (0)
- **Model:** LightGBM

```
----- CLASSIFICATION MODEL PREFOMANCE IN TEST SET-----
* R-squared model of Test: 0.9929

* Confusion Matrix of Test:
[[1557    8]
 [   5  257]]

* Classification Report of Test:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      1565
           1       0.97      0.98      0.98       262

    accuracy                           0.99      1827
   macro avg       0.98      0.99      0.99      1827
weighted avg       0.99      0.99      0.99      1827
```

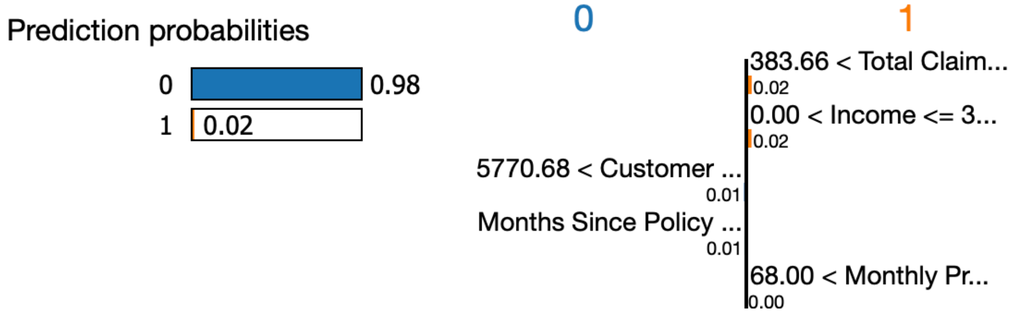
Local Model-Agnostic Methods

LIME - Tabular Data | Classification

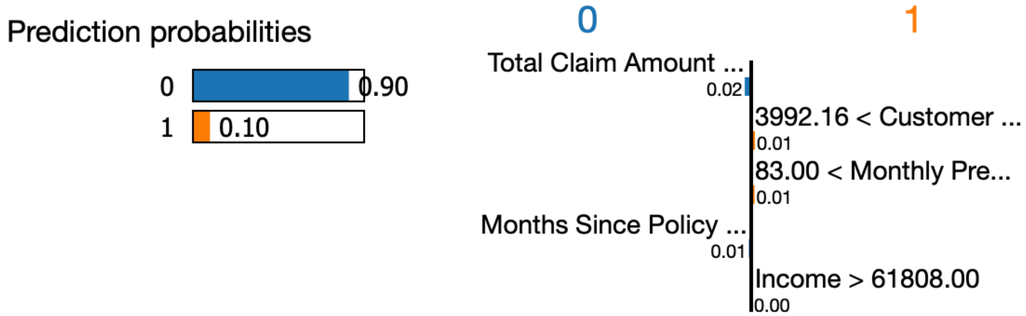
```
import lime
import lime.lime_tabular

explainer = lime.lime_tabular.LimeTabularExplainer(np.array(X_train),
                                                   feature_names = X_train.columns)

for i in range(0, 20):
    exp = explainer.explain_instance(X_test.iloc[i, :], lgb_model.predict_proba)
    exp.show_in_notebook(show_table = True)
```



Feature	Value
Income	0.00
Customer Lifetime Value	5439.34
Monthly Premium Auto	82.00
Months Since Policy Inception	40.00
Total Claim Amount	590.40



Feature	Value
Total Claim Amount	398.40
Income	32375.00
Customer Lifetime Value	6563.64
Months Since Policy Inception	83.00
Monthly Premium Auto	83.00



Feature	Value
Total Claim Amount	8.28
Customer Lifetime Value	4144.87
Monthly Premium Auto	103.00
Months Since Policy Inception	92.00
Income	64478.00

Local Model-Agnostic Methods

LIME - Text Data

	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channel! ;)	1

For	Christmas	Song	visit	my	channel!	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

Local Model-Agnostic Methods

LIME - Text Data

case	label_prob	feature	feature_weight
1	0.1701170	good	0.000000
1	0.1701170	a	0.000000
1	0.1701170	is	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	For	0.000000
2	0.9939024	;)	0.000000

Local Model-Agnostic Methods

LIME - Image

Local Model-Agnostic Methods

LIME - Local Surrogate

Advantages

- Local surrogate models benefit from the literature and experience of training and interpreting interpretable models.
- Human-friendly explanation.
- LIME is one of the few methods that works for **tabular data, text and images**.
- LIME is implemented in Python (lime library) and R (lime package and iml package) and is **very easy to use**.
- The explanations created with local surrogate models can use other (interpretable) features than the original model was trained on.

Local Model-Agnostic Methods

LIME - Local Surrogate

Disadvantages

- The correct **definition of the neighborhood is a very big, unsolved problem** when using LIME with tabular data.
- Data points are sampled from a Gaussian distribution, **ignoring the correlation between features**, can lead to unlikely data points which can then be used to learn local explanation models.
- **The instability of the explanations**, if you repeat the sampling process, then the explanations that come out can be different, it is **difficult to trust the explanations**, and you should be very critical.
- **Can be manipulated** by the data scientist to hide biases, makes it more difficult to trust explanations generated with LIME.
- *Conclusion:* Local surrogate models, with LIME as a concrete implementation, are very promising. But the method is still in development phase and many problems need to be solved before it can be safely applied.

Local Model-Agnostic Methods

Shapley value

- **Theory of Games:** assigning payout to players depending their contribution to the total.
- Shapley values was introduced in 1951 by Lloyd Shapley.
- To ML, Shapley value figure out how “payout” (feature contribution) can be distributed among features.
  - Feature: “player”
  - Prediction: “payout”

Lloyd Stowell Shapley

(1923 - 2016)

Nobel Prize (2012) in Economic Sciences

Local Model-Agnostic Methods

Shapley value | Example

Suppose you trained an ML model to predict apartment prices

You need to explain why the model predicts €300,000 for a certain apartment.

Average prediction of all apartments: €310,000.

Local Model-Agnostic Methods

Shapley value | Example

Term in Game Theory	Relation to ML	Relation to House Prices Example
Game	Prediction task for single instance of dataset	Prediction of house prices for a single instance
Gain	Actual prediction for instance - Average prediction for all instances	Prediction for house price (€300,000) - Average Prediction(€310,000) = -€10,000
Players	Feature values that contribute to prediction	‘Park=nearby’, ‘cat=banned’, ‘area=50m^2’, ‘floor=2nd’

Local Model-Agnostic Methods

Shapley value | Example

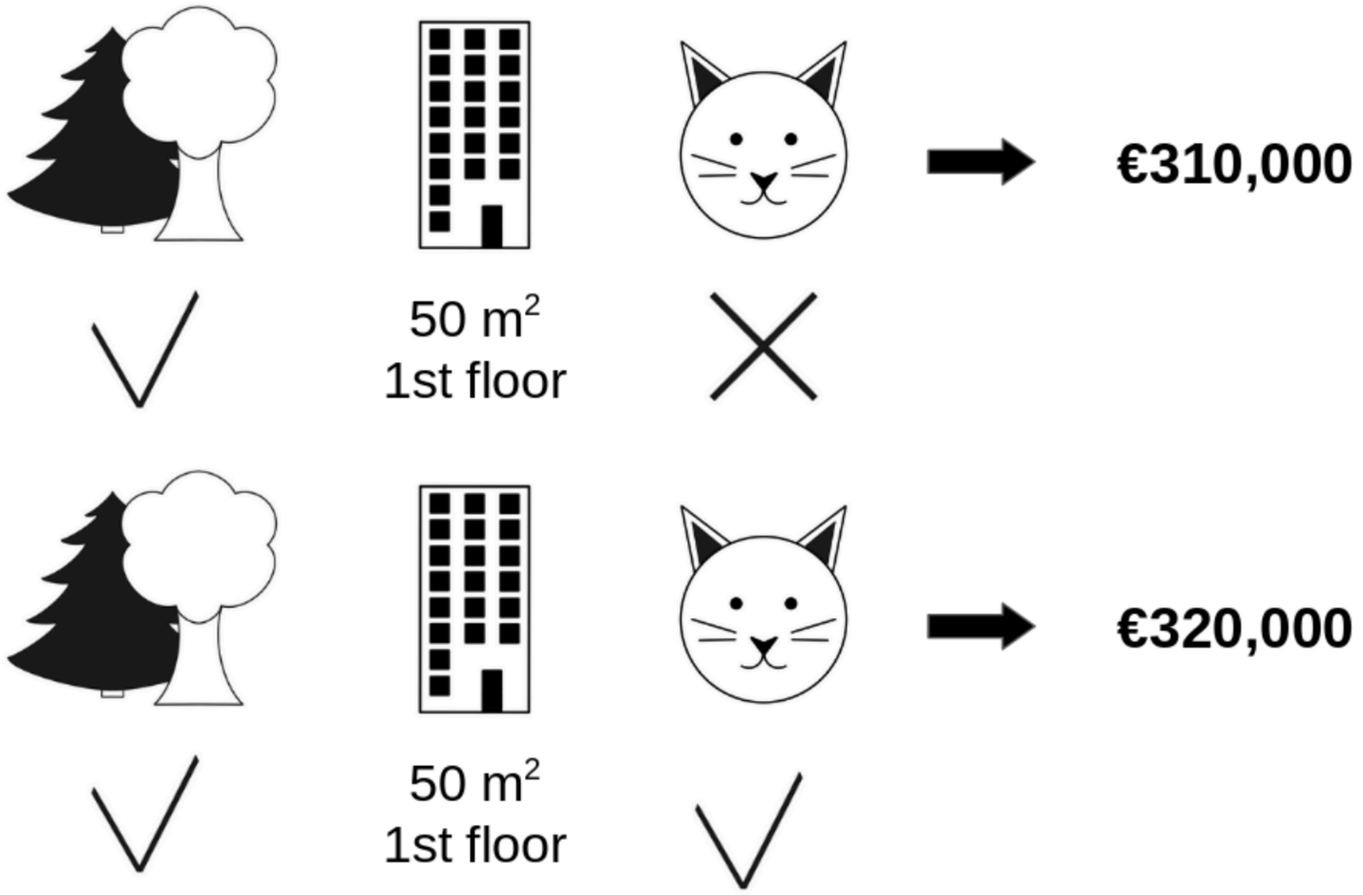
Goal :
Explain the difference between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000.

Feature	Contribution
'park-nearby'	€30,000
size-50	€10,000
floor-2nd	€0
cat-banned	-€50,000
Total: -€10,000 (Final prediction - Average Prediction)	

One possible explanation

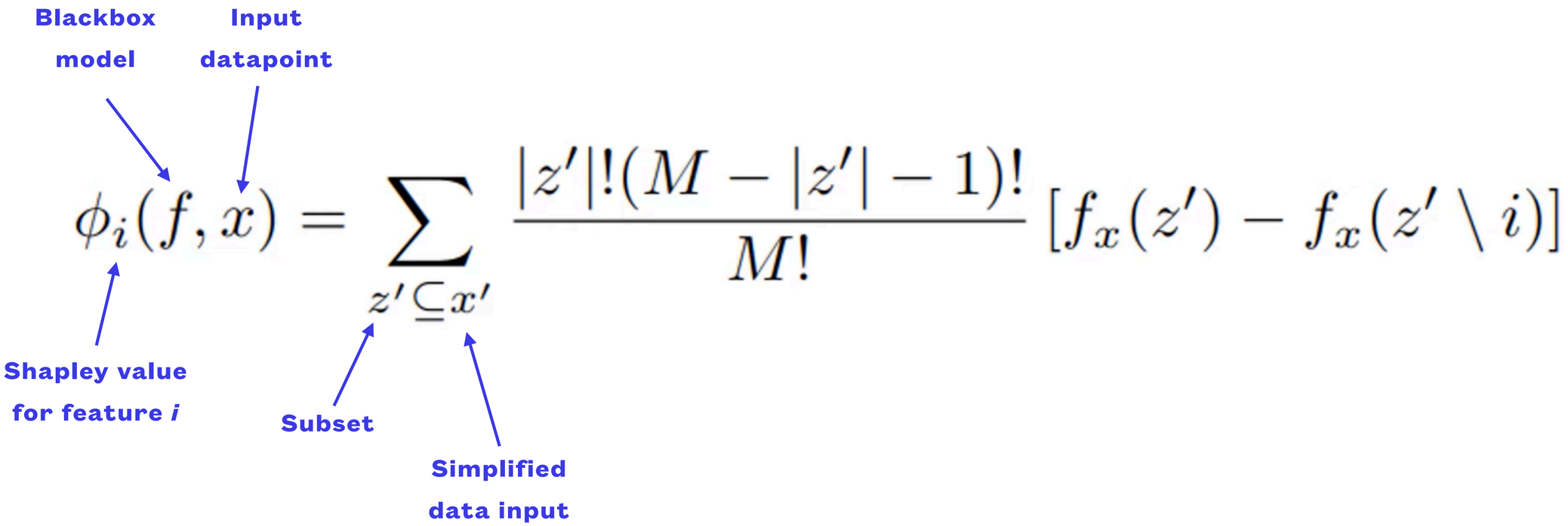
Local Model-Agnostic Methods

Shapley value | Example



Local Model-Agnostic Methods

Shapley value | Calculation



Local Model-Agnostic Methods

Shapley value | Trade off

Advantages

- **Based on solid theoretical foundation.** The axioms - Efficiency, Symmetry, Dummy, and Additivity give the explanation a reasonable foundation.
- Value is **fairly distributed** among all features, delivery full explanation.
- **Legally compliant** method in the law requires explainability situations.
- Allows **contrastive explanation** - compare to subset or single data point (that LIME do not have).

Disadvantages

- **Computationally expensive.**
- Can be easily **misinterpreted.**
- **Always use all features** to calculate explanations.
- No prediction model, so can’t be used for what if testing.
- Not work well when features are correlated.

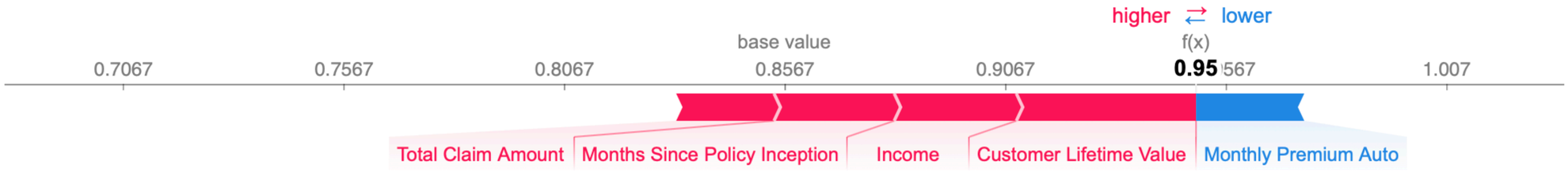
Local Model-Agnostic Methods

SHAP - SHapley Additive exPlanation

- A framework to explain individual prediction by Lundberg and Lee (2017).
- Based on the game theoretically optimal Shapley values.
- Include extension for:
  - TreeExplainer: for tree ensembles
  - KernelExaplainer: for any models by using specially-weighted local linear regression
  - DeepExplainer: for deep learning models
  - GradientExplainer: single expected value equation.

Local Model-Agnostic Methods

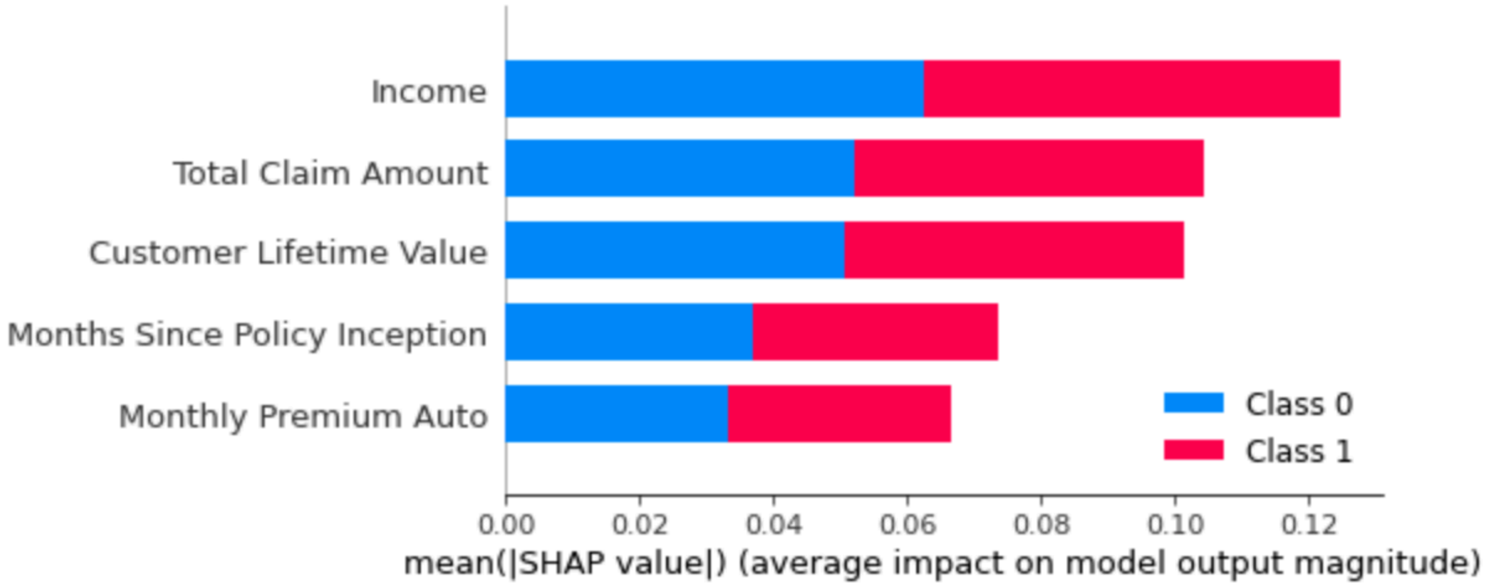
SHAP - Force Plot



- Shapley Values can be visualized as forces
- Prediction starts from the baseline (Average of all predictions)
- Each feature value is a force that increases (red) or decreases (blue) the prediction

Local Model-Agnostic Methods

SHAP - Summary Plot

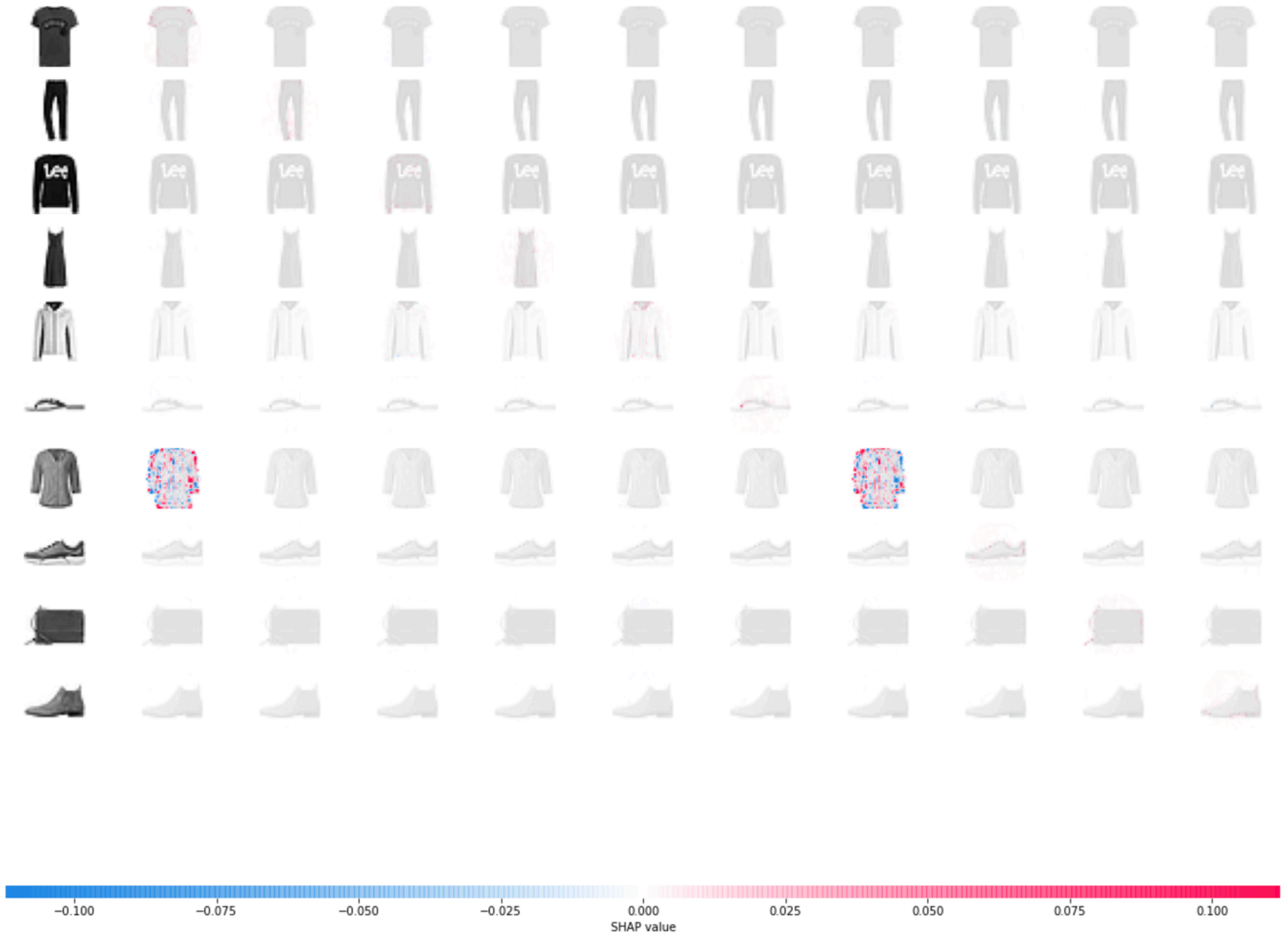


Local Model-Agnostic Methods

SHAP - Image Classification

- **Positive shap values:** red color and they represent the pixels that **contributed to classifying** that image as that particular class.

- **Negative shap values:** blue color and they represent the pixels that **contributed to NOT classify** that image as that particular class.



Local Model-Agnostic Methods

SHAP | Trade off

Advantages

- **Based on solid theoretical foundation.**
- **Fairly distributed** among all features, delivery full explanation.
- **Contrastive explanation** - compare to subset or single data point.
- **Connect LIME & Shapley values.**
- **Fast implementation for tree-based models.**
- **Global model interpretations.**

Disadvantages

- **KernelSHAP is slow.**
- **KernelSHAP ignores feature dependence.**
- **TreeSHAP** can produce **unintuitive feature attributions**.
- Can create intentionally misleading interpretations.

The Future of Interpretability

- The **focus** will be on **model-agnostic interpretability tools**.
- Machine learning will be **automated** and, with it, interpretability.
- We do not analyze data, we **analyze models**.
- The data scientists will automate themselves.
- Robots and programs will **explain themselves**.
- Interpretability could **boost machine intelligence research.**

References

"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

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the "trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the "trusting the model" problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model.

LIME

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

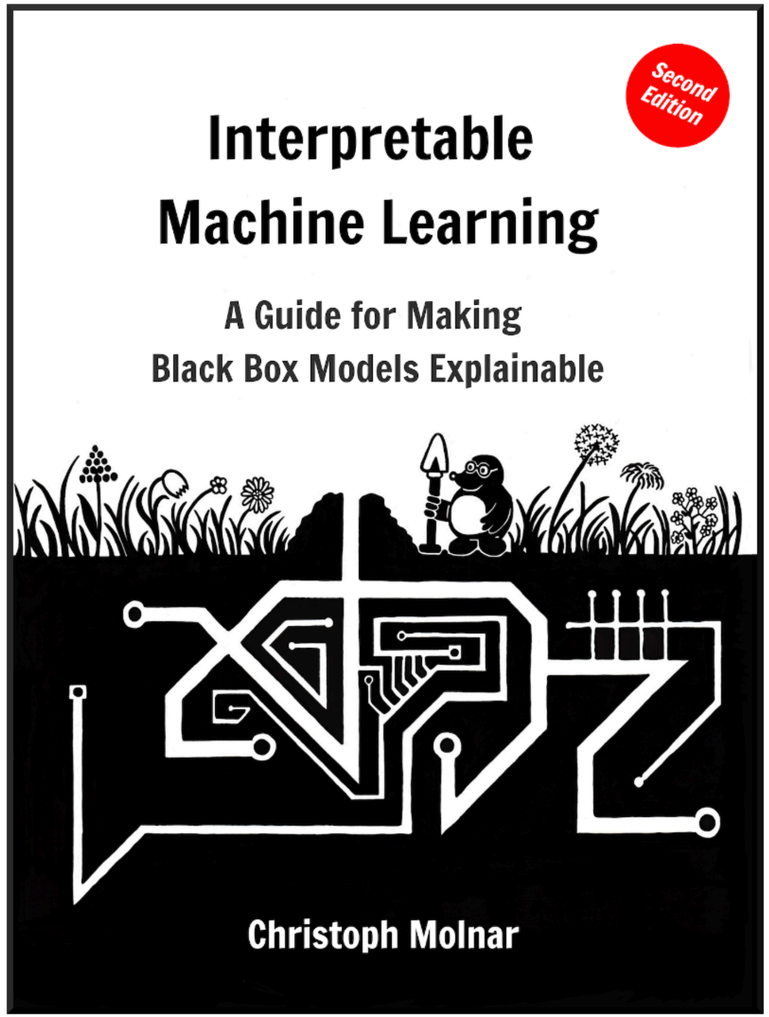
Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

SHAP

References


Second Edition
Interpretable Machine Learning
A Guide for Making Black Box Models Explainable
Christoph Molnar

References

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)

Alvarez-Melis, David, and Tommi S. Jaakkola. "On the robustness of interpretability methods." arXiv preprint arXiv:1806.08049 (2018)

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186 (2020)

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019)

Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." International Conference on Artificial Intelligence and Statistics. PMLR (2020)

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186 (2020)

Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317

Štrumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

Thank you!