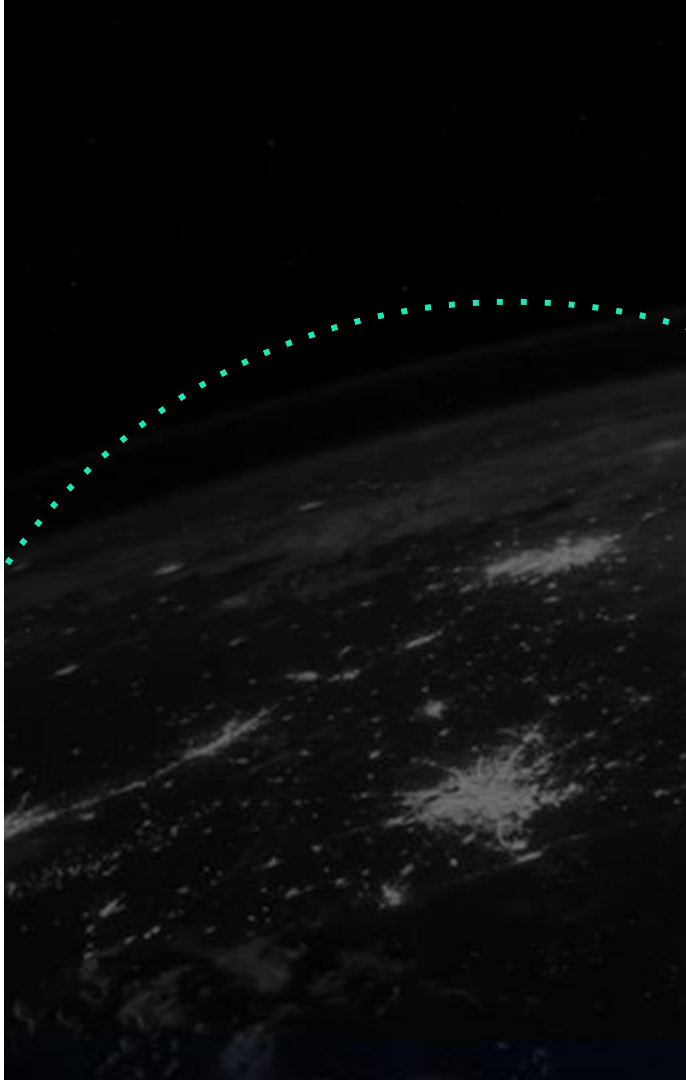
SIAPARTNERS

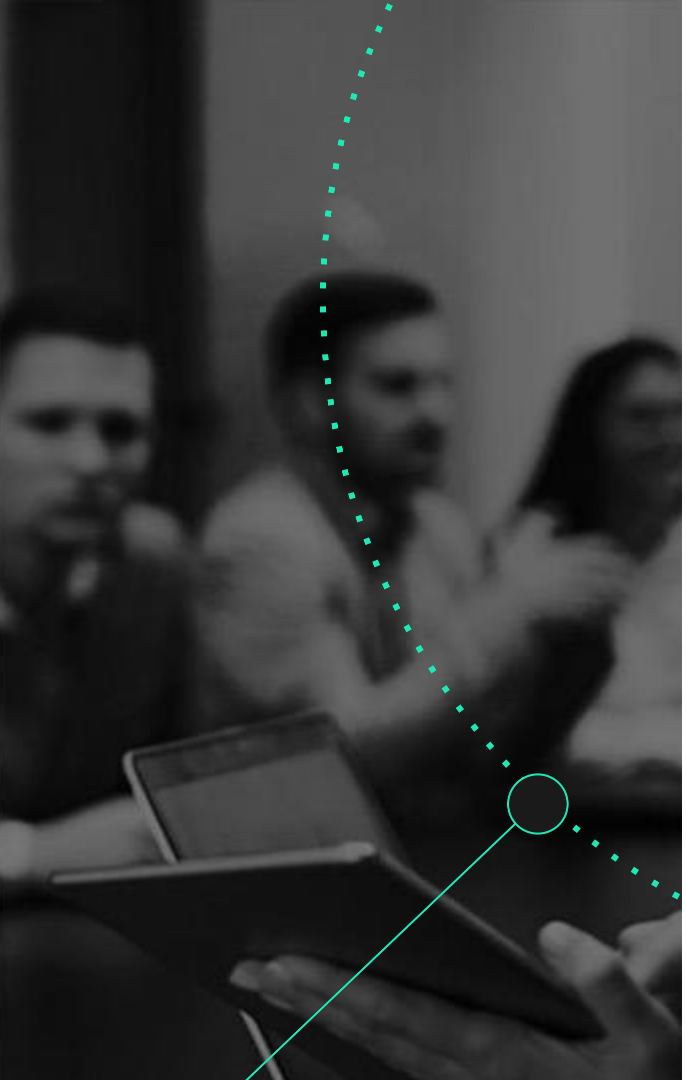
2023 – Baise

Important skills for practical machine learning applications

Loïc Roldán Waals & Reinier Kruisbrink

loic.roldanwaals@sia-partners.com
reinier.kruisbrink@sia-partners.com





Today’s outline

1. Introduction
2. Use the Right Metrics
3. Explain Your Model’s Decisions
4. Q&A

1.
Introduction

We are a next-generation consulting firm

We are a global firm that has grown steadily over the past 20 years

2,000 Consultants

36 Offices across **18** countries

390M$ in revenue for FY21/22

We invest heavily in tech and design to stay on cutting-edge and meet our clients' evolving challenges

5 AI centers

2 Design Centers

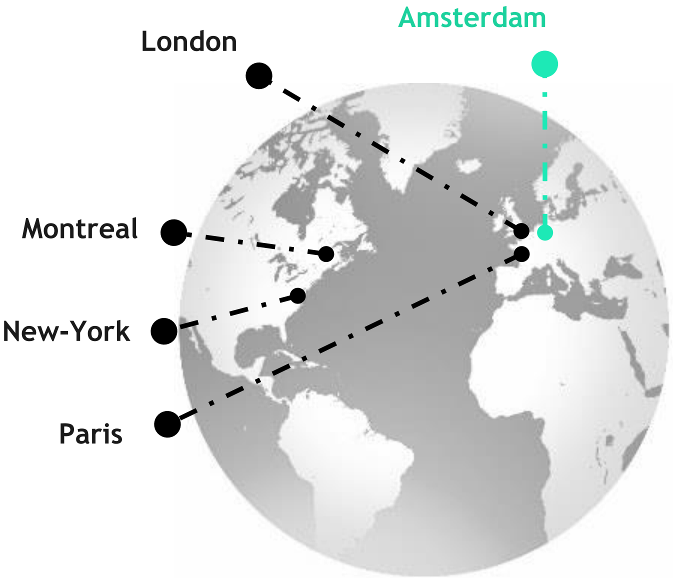
600 Clients
92% returning

We cultivate expertise stemming from R&D activities and our proximity with our clients' industries

4% Of our revenue invested in R&D

130k+ Followers on LinkedIn

We are a global Data Science business unit, with 5 centers of excellence



150
Data scientists, Web Developers, Data Engineers, UI/UX designers

5
AI Centers of Excellence

31
Offices with AI ambassadors

+50 clients
Trust us

Heka
Our AI ecosystem

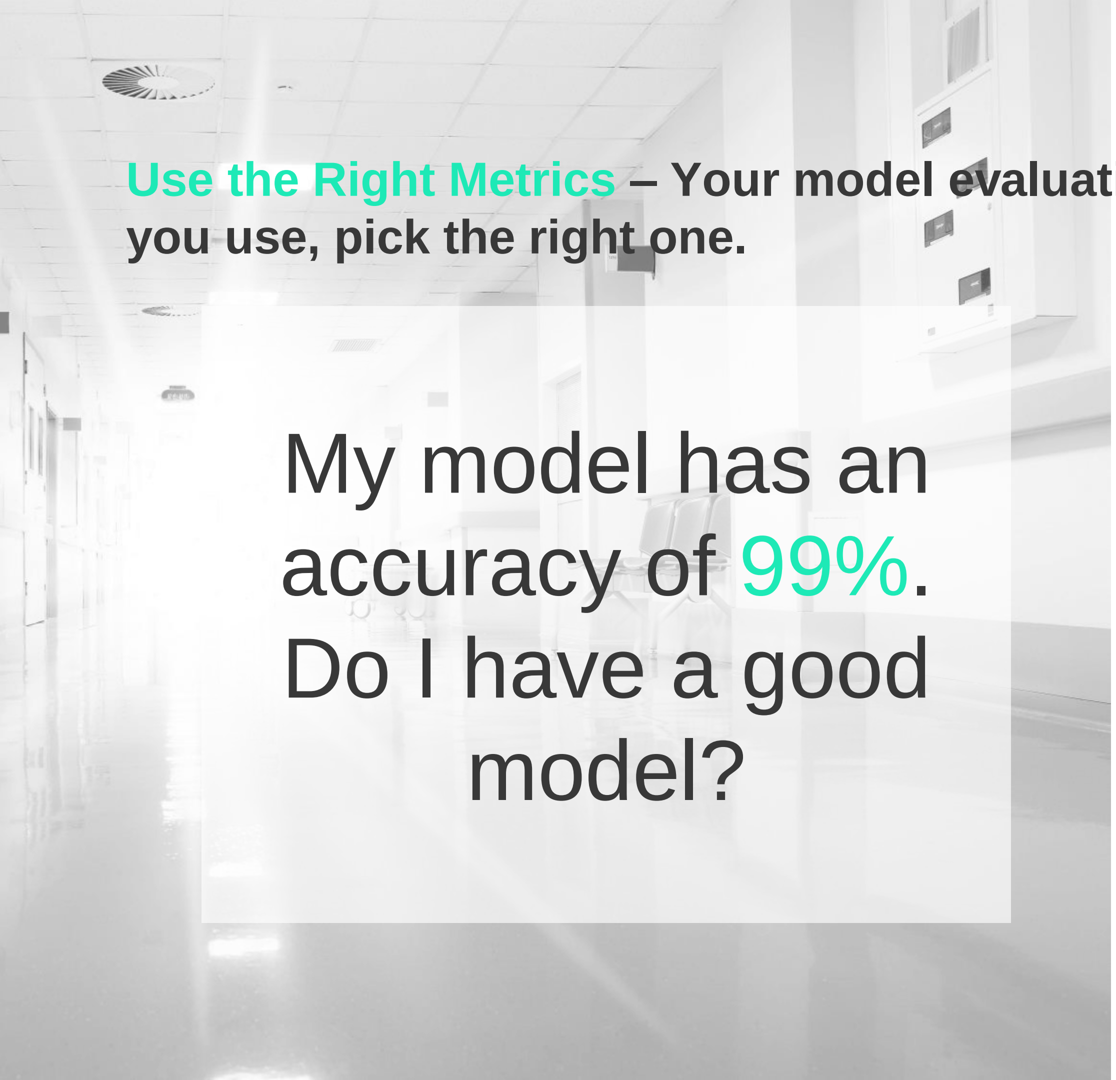
Ready-to-use AI solutions
Customer experience, operational efficiency, targeted offerings, etc.

AI accelerators
Catalog of 70+ AI building blocks to accelerate ideation and development of AI projects

Platform As A Service
Admin features, production & development environments

2.

Use the Right Metrics



Use the Right Metrics **– Your model evaluation is shaped by the metric you use, pick the right one.**

My model has an accuracy of 99%. Do I have a good model?

Use the Right Metrics – Your model evaluation is shaped by the metric you use, pick the right one.

My model has an accuracy of 99%. Do I have a good model?

It depends on distribution of the target variable

Use the Right Metrics – Each metric has it’s pros and cons and a specific situation where it should be used.

We will discuss 3 regression metrics.

Adjusted R^2	RMSE	MAE
Pro: • Easily comparable between regression problems (0-1) Con: • Not easy to quantify the final impact of your model on business case.	Pro: • Unit is on the same level as the prediction error. • Used by many models to optimise the fit. Con: • Sensitive to outliers (so use this when they should be weighted extra). • Not easy to compare across regression problems.	Pro: • Easiest to interpret. • Easy to measure impact on business case. Con: • Hard to numerically optimise, so could lead to you optimising on one metric while presenting another. • Not easy to compare across regression problems.

Use the Right Metrics – Each metric has it’s pros and cons and a specific situation where it should be used.

We will discuss 3 classification metrics.

Accuracy

Pro:
- Easy to interpret and explain.

Con:
- Not suitable for imbalanced datasets.

ROC AUC & PR AUC

Pro:
- Encompasses all thresholds.
- ROC useful when model outputs are used for ranking.
- PR curve useful for imbalanced data.

Con:
- More abstract.
- More difficult to explain to stakeholders.

F-Score

Pro:
- Easier to explain than AUC.
- Can handle imbalanced data.

Con:
- Need to determine preference between precision and recall

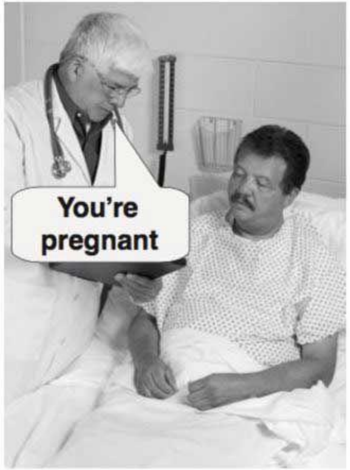
Use the Right Metrics – Familiarise yourself with the confusion matrix and use accuracy when you have balanced, equally important classes.

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

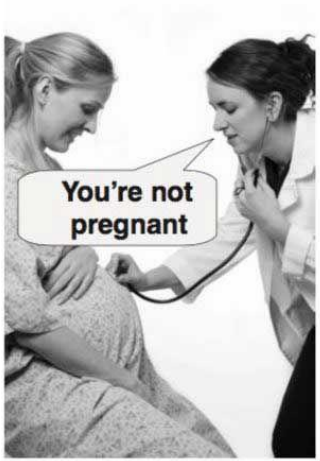
Confusion Matrix Overview

Use the Right Metrics – Familiarise yourself with the confusion matrix and use accuracy when you have balanced, equally important classes.

Type I error
(false positive)



Type II error
(false negative)



Predicted Class

	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Confusion Matrix Overview

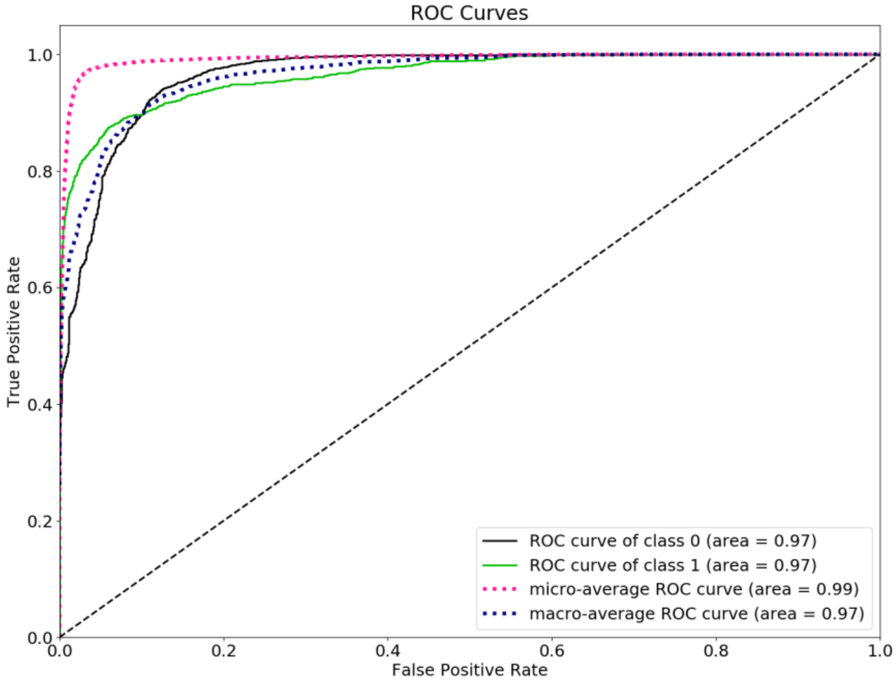
Use the Right Metrics – Changing the threshold will change your predictions and can also change your metrics.

ID	Actual	Prediction Probability	>0.6	>0.7	> 0.8	Metric
1	0	0.98	1	1	1	
2	1	0.67	1	0	0	
3	1	0.58	0	0	0	
4	0	0.78	1	1	0	
5	1	0.85	1	1	1	
6	0	0.86	1	1	1	
7	0	0.79	1	1	0	
8	0	0.89	1	1	1	
9	1	0.82	1	1	1	
10	0	0.86	1	1	1	
			0.75	0.5	0.5	TPR
			1	1	0.66	FPR
			0	0	0.33	TNR
			0.25	0.5	0.5	FNR

Example of how thresholds can affect metrics

Use the Right Metrics – The ROC curve plots the tradeoff between the false positive rate and the true positive rate.

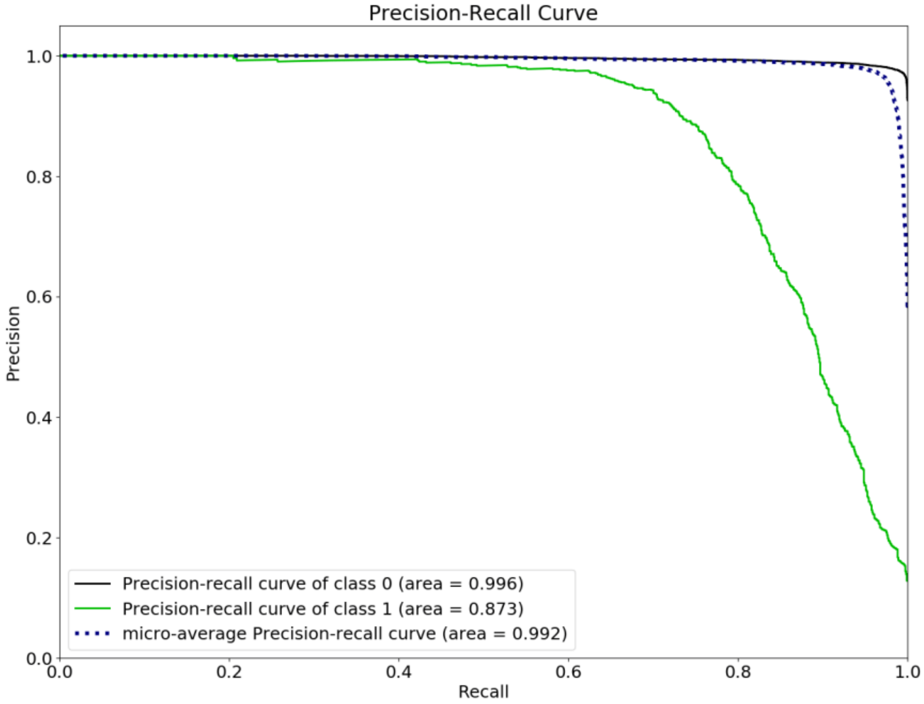
- ROC stands for "receiver operating characteristic"
- FP rate = 1 – specificity
- TP rate = sensitivity
- Requires probability values to compute.
- Checks the relationship between FPR and TPR over all thresholds.
- The larger the area under this curve (AUC) the better.
- The ROC AUC tells us how good at ranking predictions our model is.
- Use when you care equally about classes.



ROC Curve Example

Use the Right Metrics – The PR curve plots the tradeoff between the precision and recall.

- PR stands for “Precision Recall”
- Recall = sensitivity
- Requires probability values to compute.
- The larger AUC the better.
- The PR AUC tells us the precision and recall scores calculated for each threshold.
- Use when classes are imbalanced and you care more about the positive class.



PR Curve Example

Use the Right Metrics – The PR curve plots the tradeoff between the precision and recall.

- To choose a threshold we need to determine the costs/rewards for the confusion matrix on the right, given our business problem.
- We then choose a threshold that maximizes the formula below.
- For our convenience we can plot this as a line graph with the reward on the y-axis and the threshold on the x-axis.

	Predicted wildfire	Predicted no wildfire
Actually wildfire	The benefit of successfully predicting a wildfire (tpb).	Cost of missing a wildfire (fnc).
Actually no wildfire	The cost of thinking there will be a wildfire when there isn't one (fpc)	The benefit of successfully identifying that there is no wildfire (tnb)

$$total\ reward = \#TP * tpb + \#TN * tnb - \#FP * fpc - \#FN * fnc$$

Use the Right Metrics – F-scores are a harmonic mean between precision and recall, decide with the stakeholder how much each is worth.

$$F_{beta} = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall}$$

F-score formula

- Recall = sensitivity
- Requires probability values to compute.
- Choose the threshold that provides the highest F-score
- You can specify how much you care between precision and recall (F1 = equally important).
- Use when classes are imbalanced and you care more about the positive class.
- Slightly easier to explain to the business.

3.

Explain Your Model's Decisions

Explain Your Model's Decisions – Explaining why your model made a decision can be valuable and is sometimes necessary.

- Legal reasons/right to understand decisions affecting oneself requires an explanation on why the model made a decision.
- Can help you identify why your model is making weird predictions.
- Builds trust and provide ethical justifications
- The way in which the model makes decisions might be an insight itself.

Explain Your Model's Decisions – There are many methods to explain a model's behaviour.

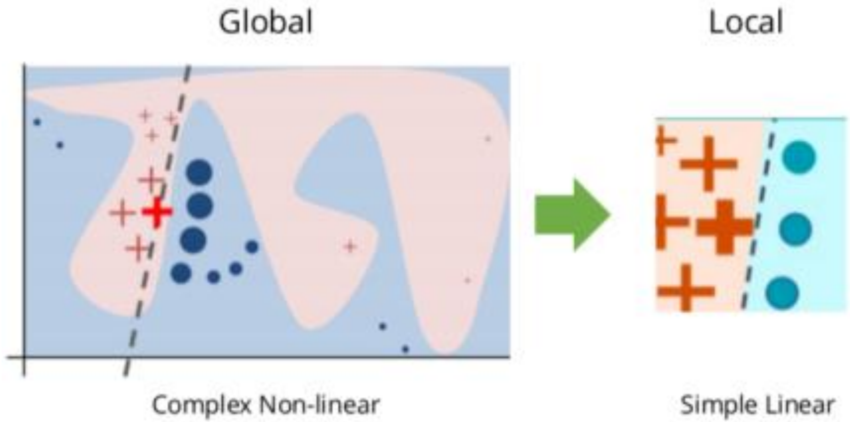
- Some models have it built-in (regression coefficients)
- Some have approximants (Feature importance in random forrests and activations in Neural Networks)
- Two general methods:
  - Local Interpretable Model-Agnostic Explanations (for single predictions)
  - SHapley Additive exPlanations (for general or single predictions)

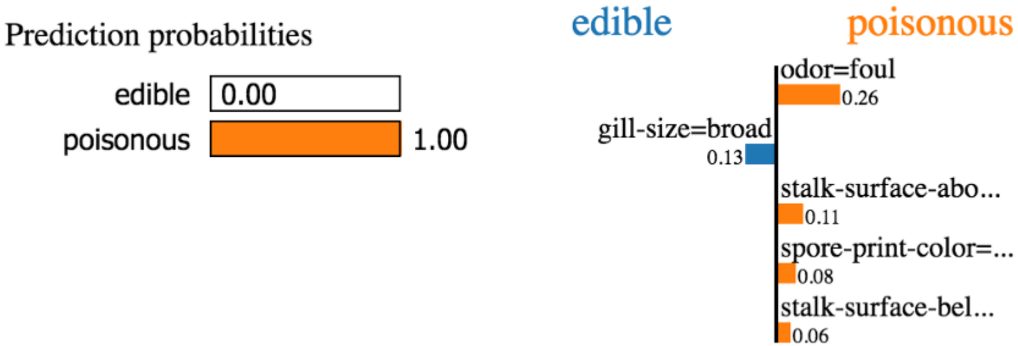
LIME

SHAP

Explain Your Model's Decisions – LIME only looks at the local space and fits a small linear model to explain the decisions.

Lime is only an approximation and only works on a single prediction. Be careful not to generalise.

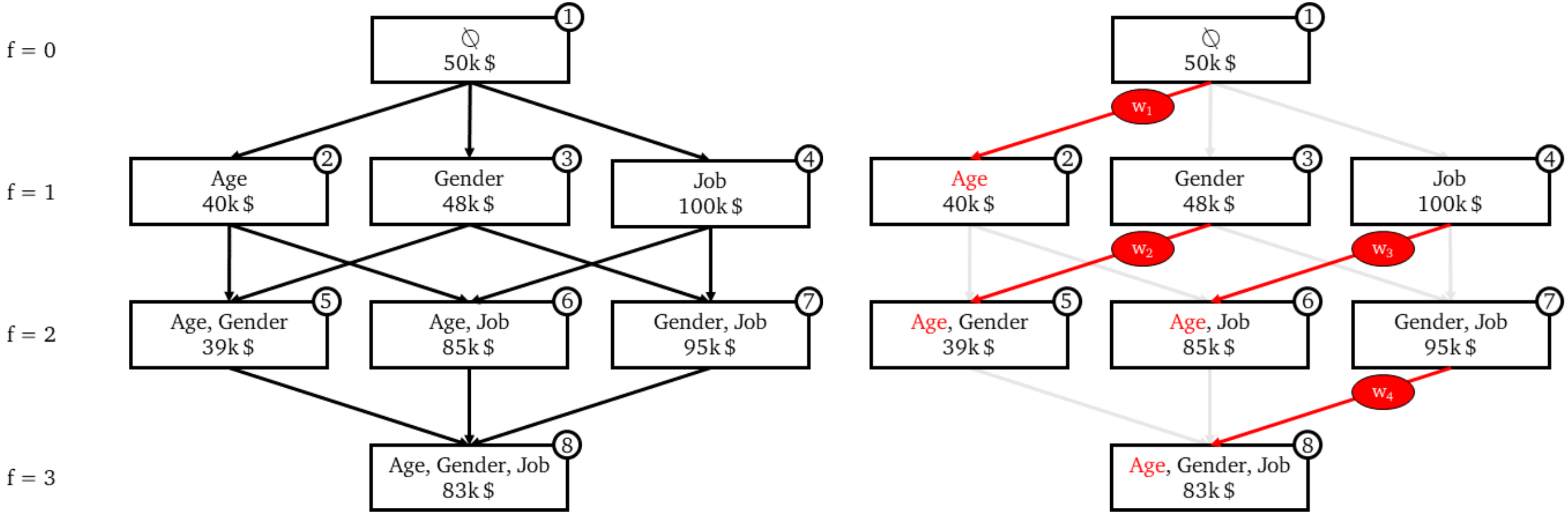




Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

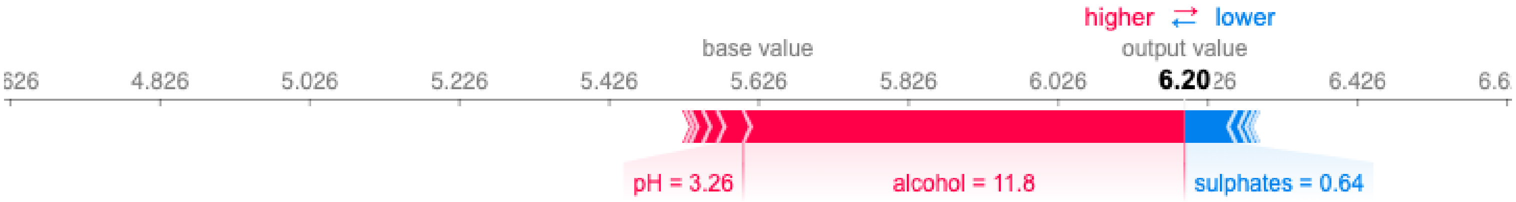
LIME output example

Explain Your Model's Decisions – SHAP uses game theory to determine the effect of each feature.



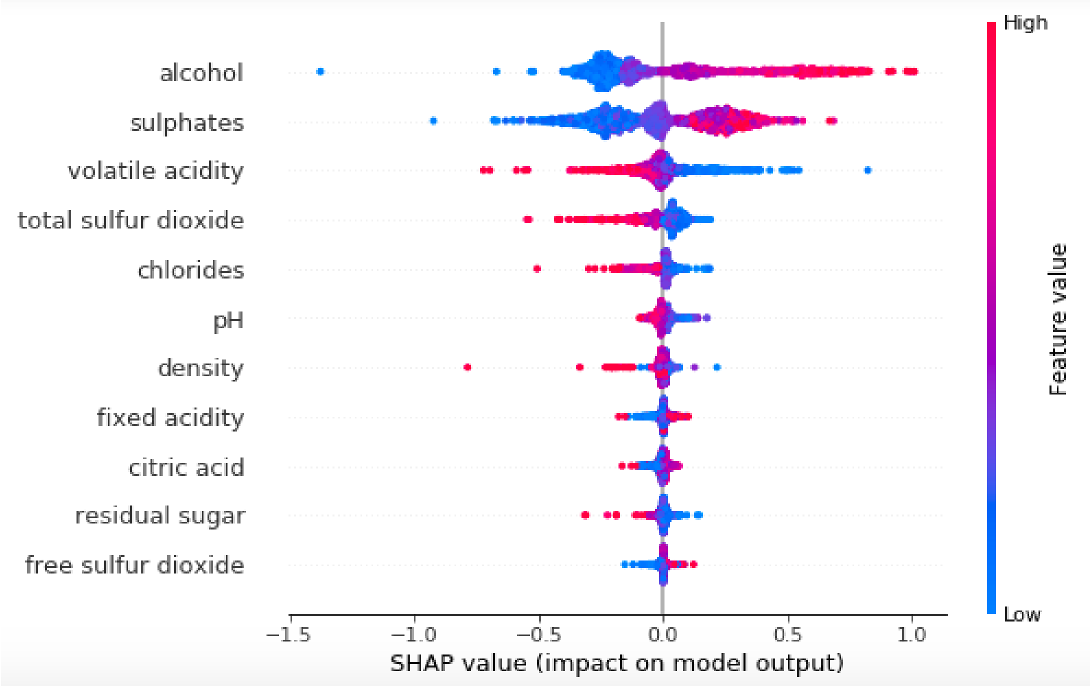
Explain Your Model's Decisions – SHAP uses game theory to determine the effect of each feature, this can be used to explain single predictions.

- SHAP values provide both instance and model interpretability.
- SHAP has very strong theoretical foundation.
- SHAP works by combining the effects of adding a given feature one at a time (in all possible combinations).
- To calculate SHAP values you need to train $2^{\#\ of\ features}$ models. So models with many features become infeasible.



SHAP single instance explanation output example

Explain Your Model's Decisions – SHAP uses game theory to determine the effect of each feature, this can be used to explain the entire model.



SHAP model explanation output example

4.

Questions

Sources

Juicero - https://www.bbc.com/news/business-39664483

Questions to help explore an data science problem
https://www.datascience-pm.com/10-questions-to-ask-before-starting-a-data-science-project/

Structured problem solving
https://strategyu.co/mckinsey-structured-problem-solving-secrets/

Time spent on tasks by data scientists
https://www.anaconda.com/state-of-data-science-2020

Overview between accuracy, roc auc, pr auc and f scores
https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Sources

ROC considerations
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.9777&rep=rep1&type=pdf

AUC PR vs AUC ROC
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/

LIME vs SHAP
https://towardsdatascience.com/lime-vs-shap-which-is-better-for-explaining-machine-learning-models-d68d8290bb16

Explanation of how LIME works
https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b
https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5

Explanation of how SHAP works
https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30
https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Sources

Calculating Adjusted R squared
https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/adjusted-r2/

Over view of 3 regression metrics
https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b