A World Leading SFI Research Centre

MLGIG Team - Milk Lactose Prediction Data Challenge - P2: Explain the models




Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

[georgiana.ifrim   thach.lenguyen timi.aderinola davide.serramazza] @insight-centre.org

4th International Workshop on Spectroscopy and Chemometrics 2024
30/04/2024

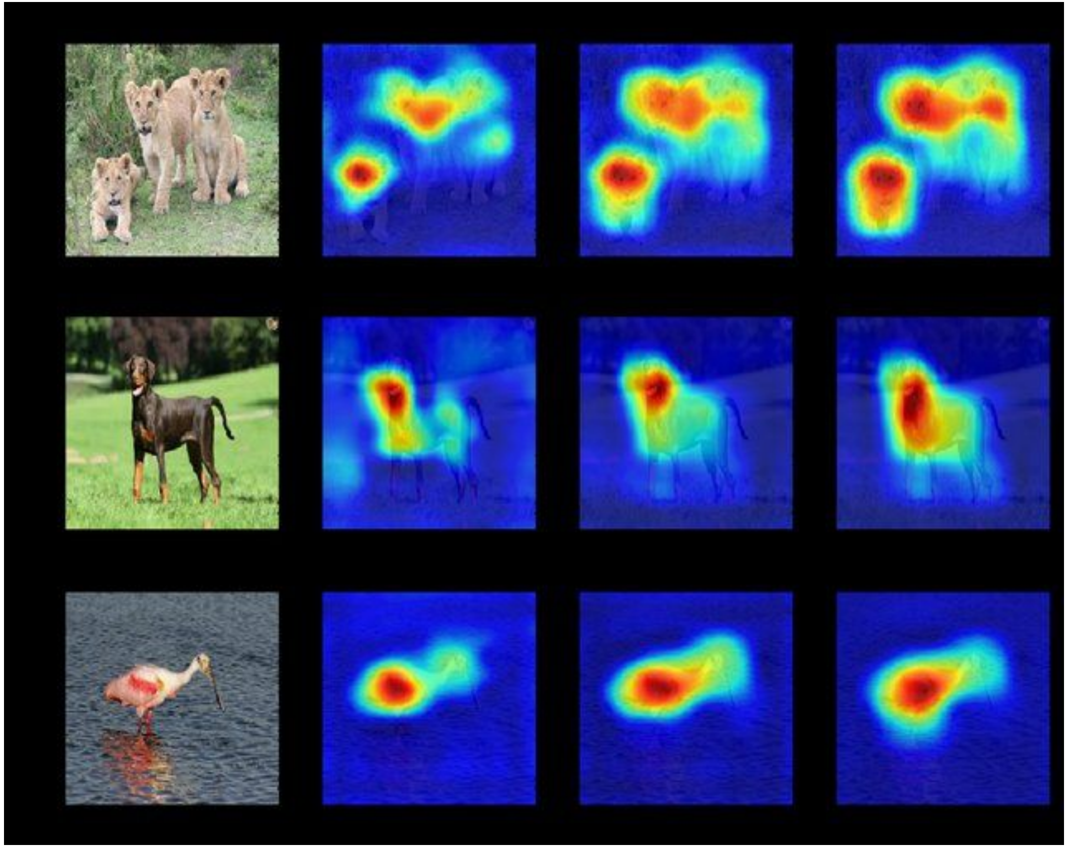
HOST INSTITUTIONS

PARTNER INSTITUTIONS

FUNDED BY:

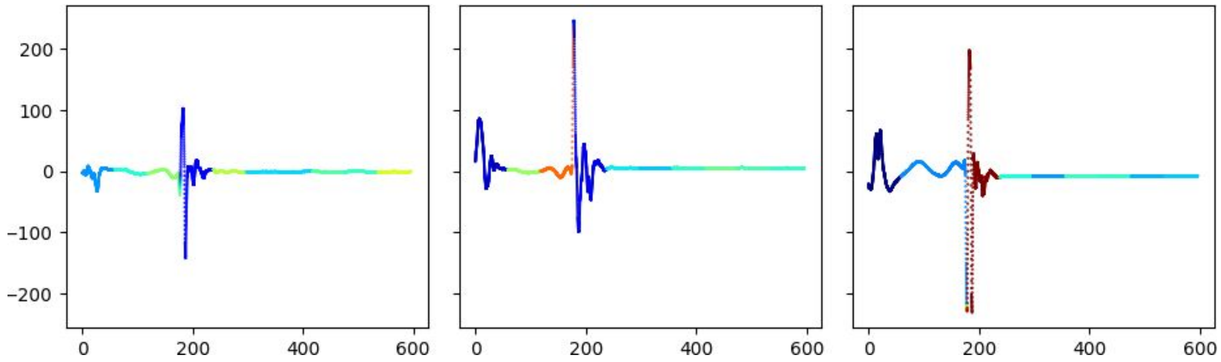
Our motivation

- We are interested in **WHY** ? E.g., why the model predicted 24 but not 60 ? on which part of the data the model was focusing ?
- We don’t have the necessary background to understand the data but can we learn something from the model ? E.g. which area is more important, which is noisy.
- We want explanation so:
  - Perhaps we can improve the model performance.
  - We can take the explanation to the domain experts.
- We have experience with time series **saliency map / attribution methods** so we used what we know.



*https://www.geeksforgeeks.org/what-is-saliency-map/

An example of saliency map for image classification.



An example of saliency map for time series classification.

Time Series Attributions

0.4	1	5	2
0.3	2	4	3
0.1	2	2	6

Sample T

$\xrightarrow[M]{D}$

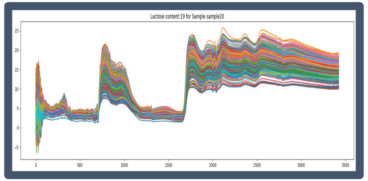
0.0	0.0	0.8	0.4
0.0	0.1	0.7	0.3
0.1	0.1	0.6	0.4

Attribution A

- Attribution (saliency map) A has the **same shape** as input T.
- Each entry in A "**explain**" the corresponding entry in T ,e.g., how important it is to the model M when it makes the prediction.
- **A = D(T,M)** where D is an attribution method.

How attribution methods work

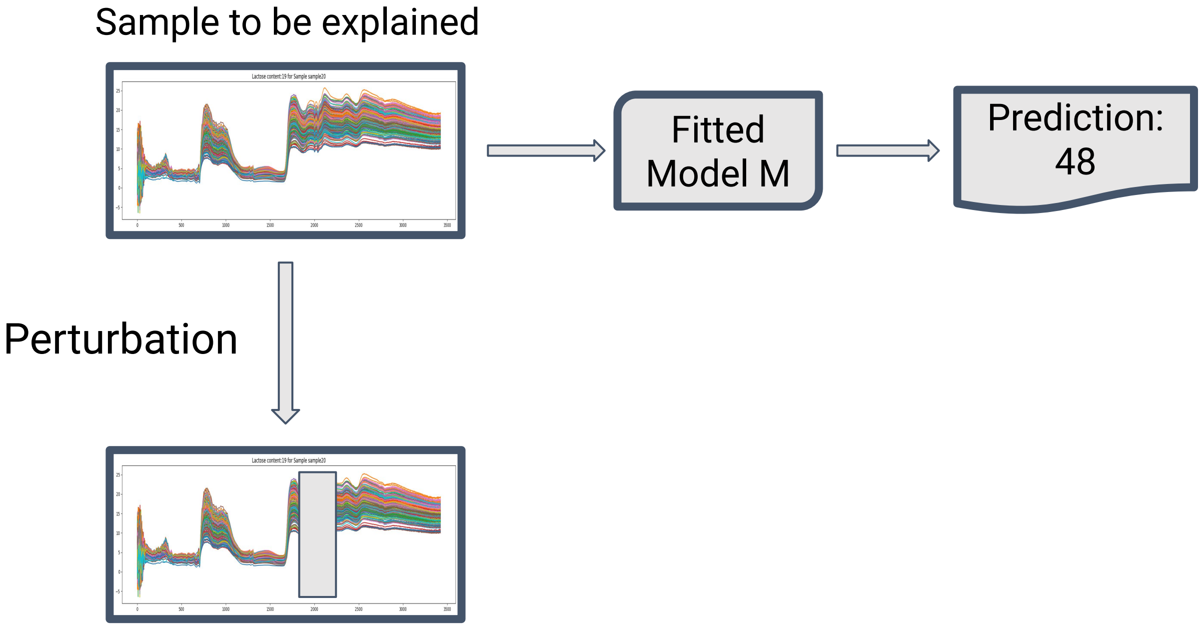
Sample to be explained



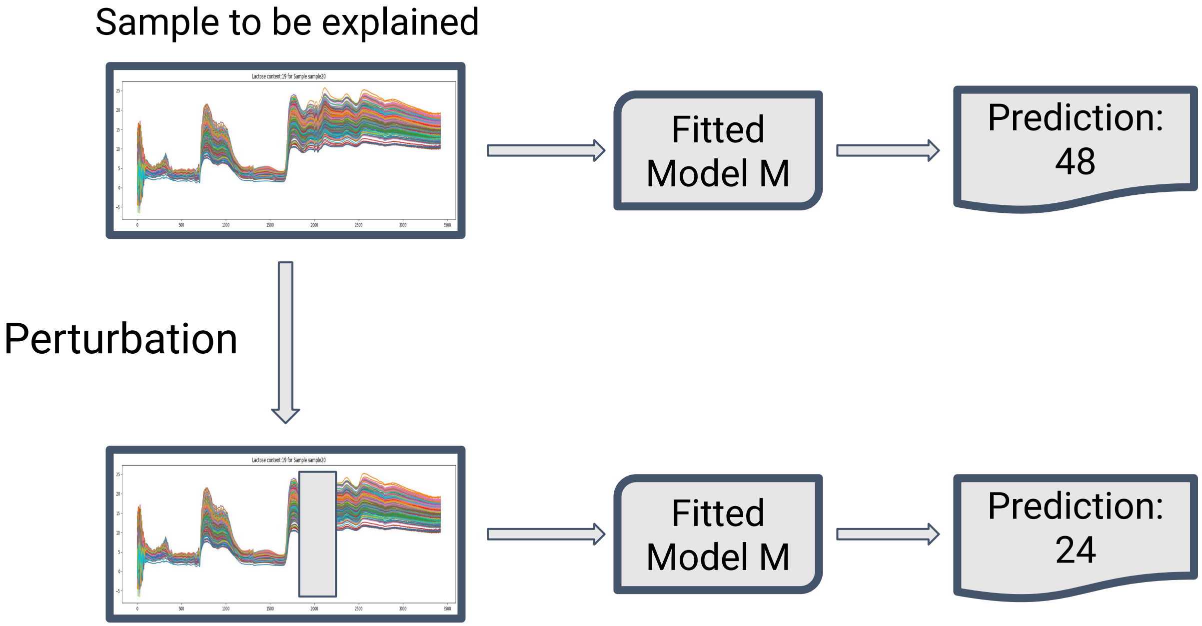
Fitted Model M

Prediction: 48

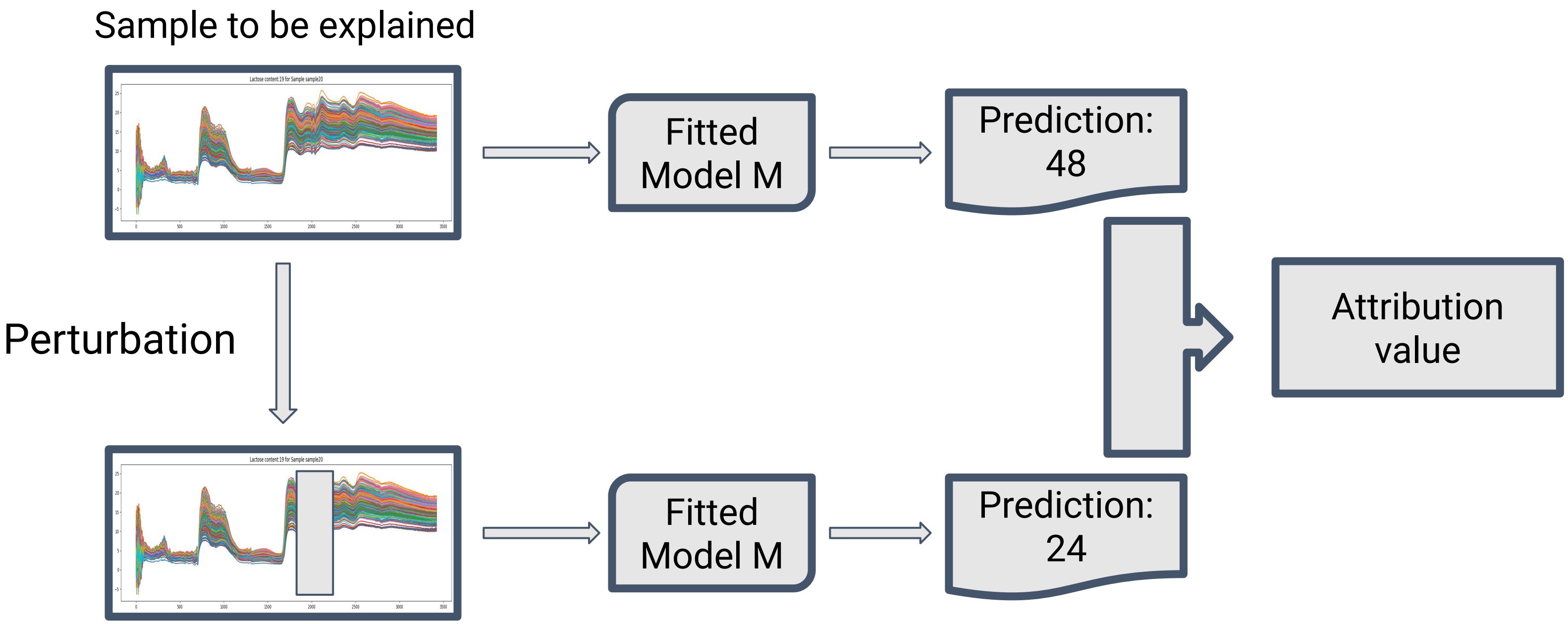
How attribution methods work



How attribution methods work



How attribution methods work



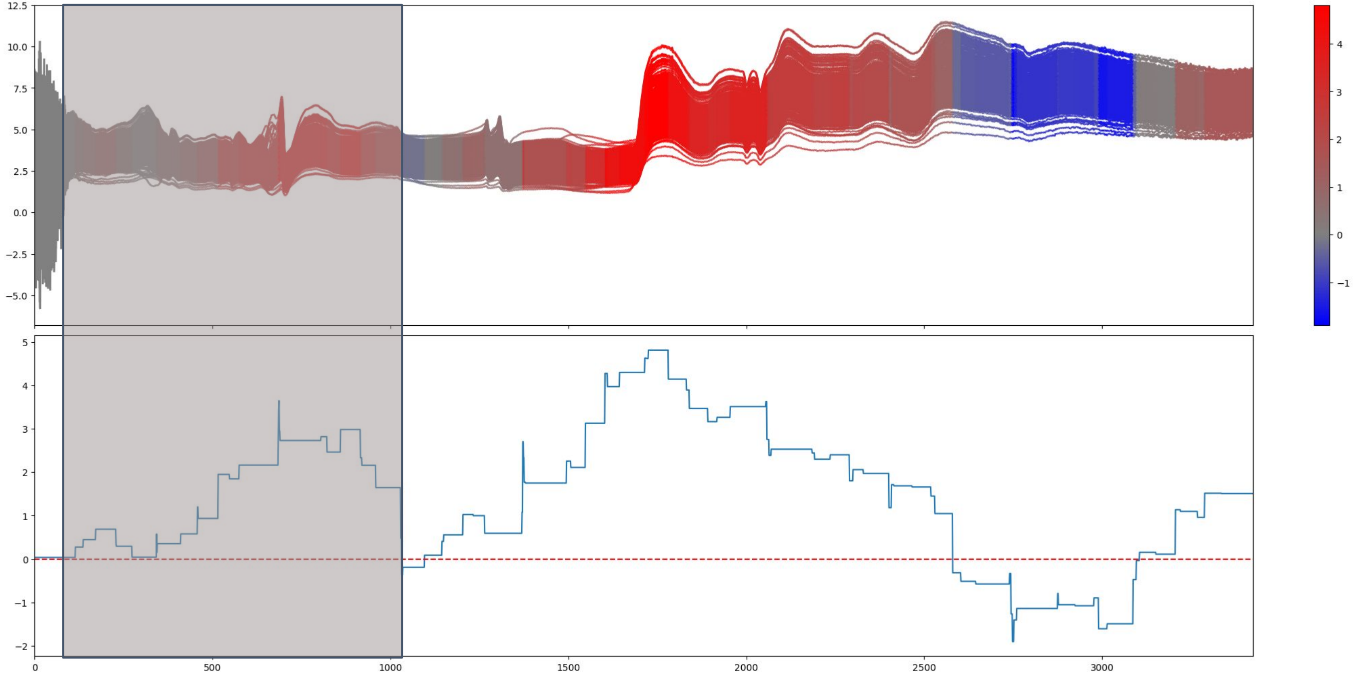
Perturbation-based **Attribution Methods**

- Post-hoc methods: i.e., they work with any black box models.
- In our experience with time series models, Shapley Value or Feature Ablation are the good options to try.
- They can be computationally expensive (~1 hour to compute the attribution of all samples using Shapley).
- We use Captum (captum.ai) implementation.
- Our **best model** + **Shapley** => attribution profile

Results

T

A



Sample 16

True Lactose: 48

Prediction: 48

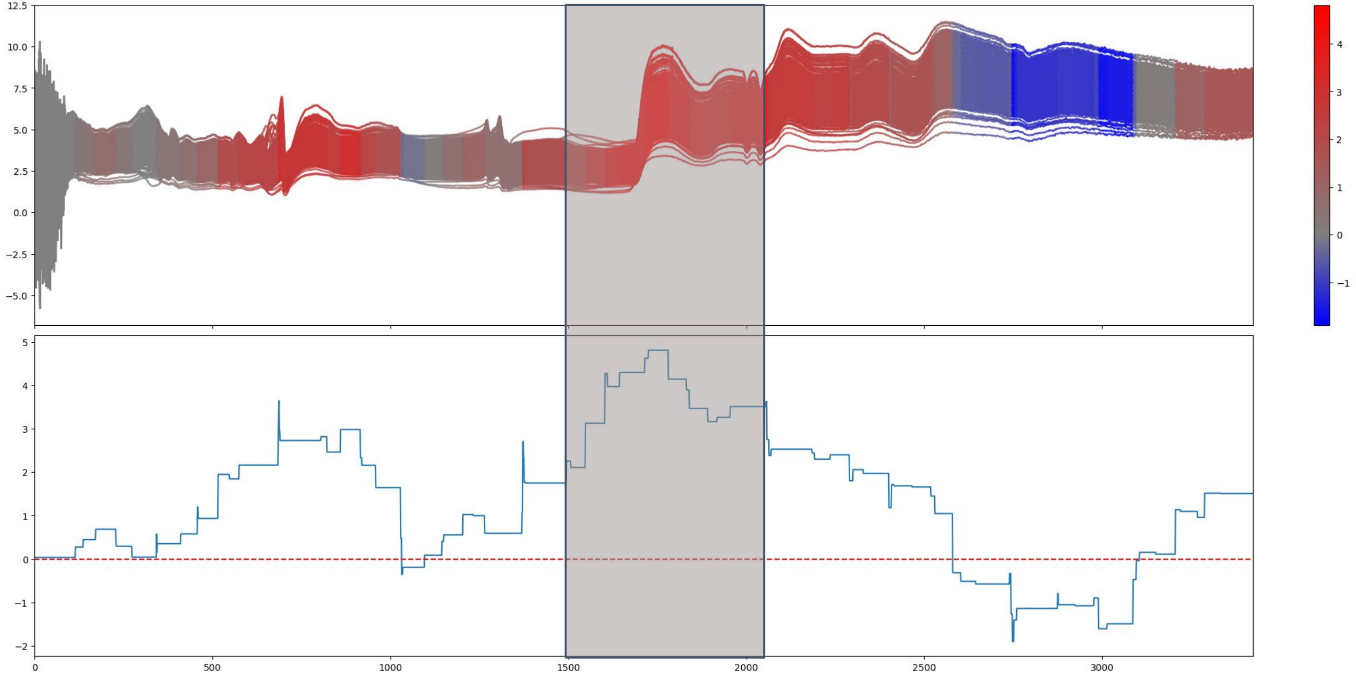
New prediction: 42.275

Insight

Results

T

A



Sample 16

True Lactose: 48

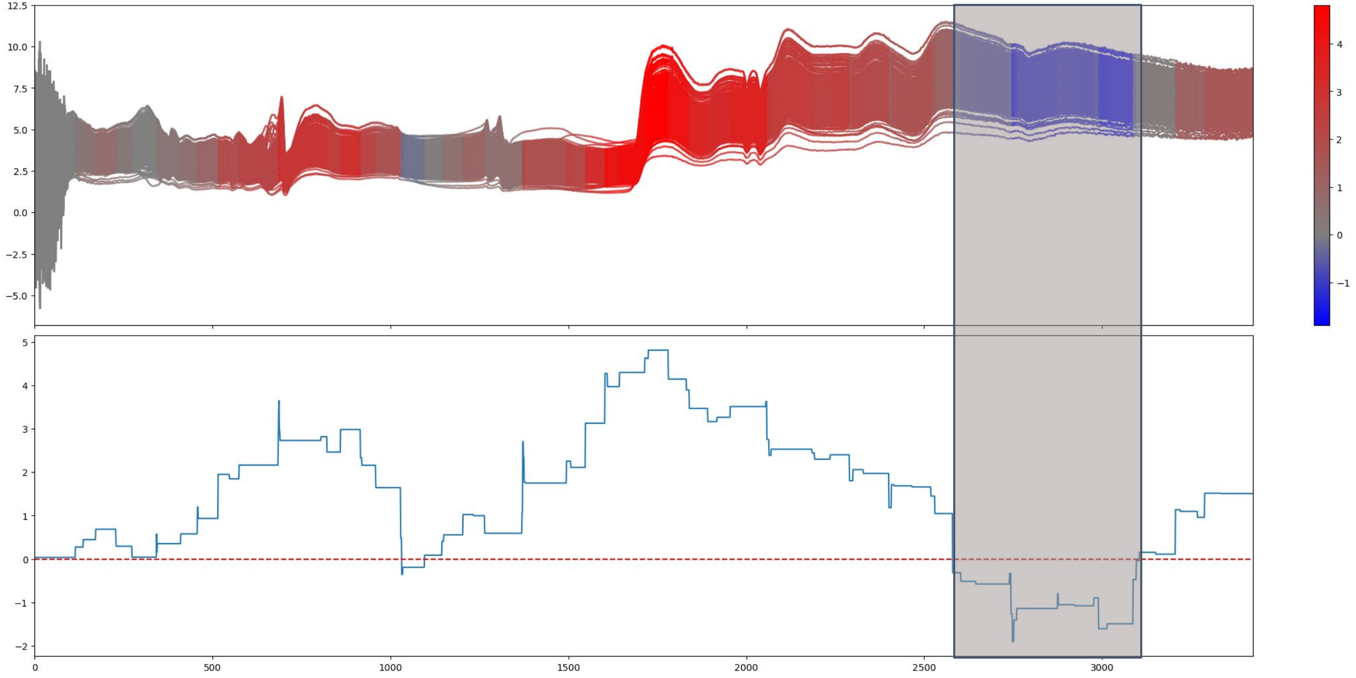
Prediction: 48

New prediction: 25.64

Results

T

A



Sample 16

True Lactose: 48

Prediction: 48

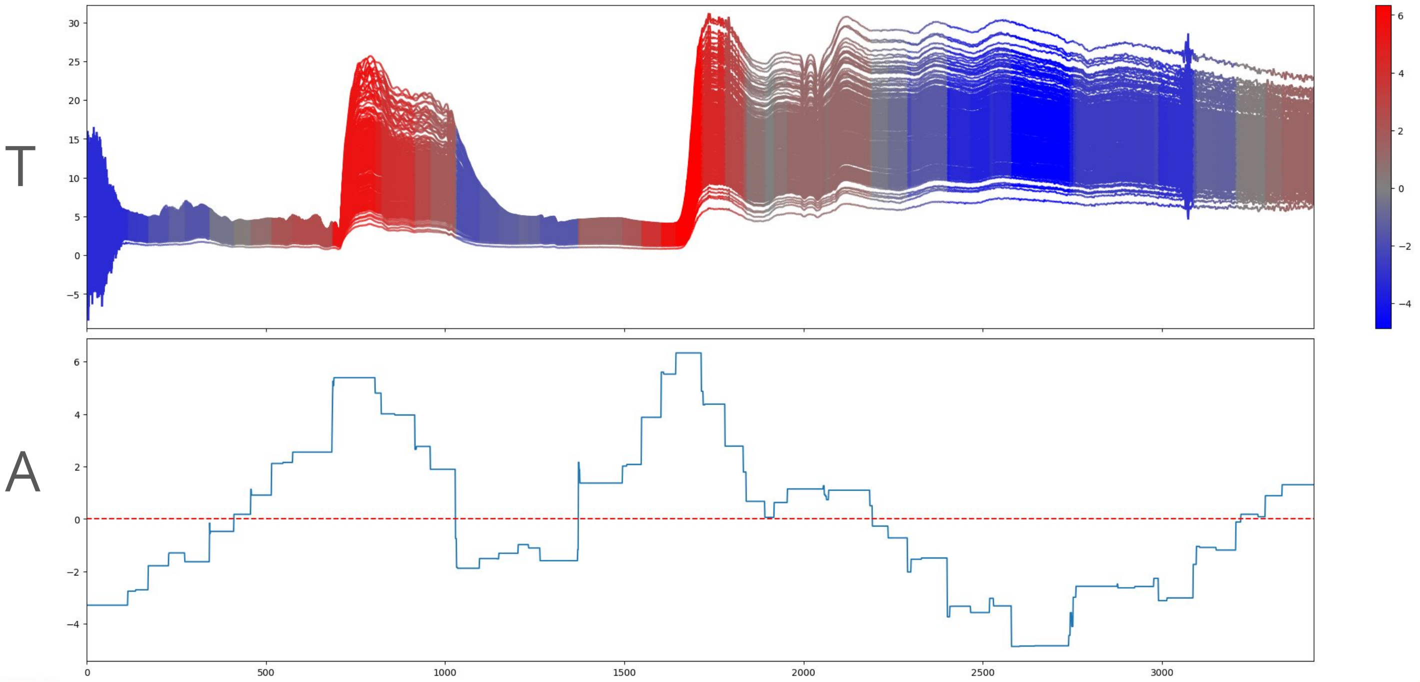
New prediction: 61.035

Insight

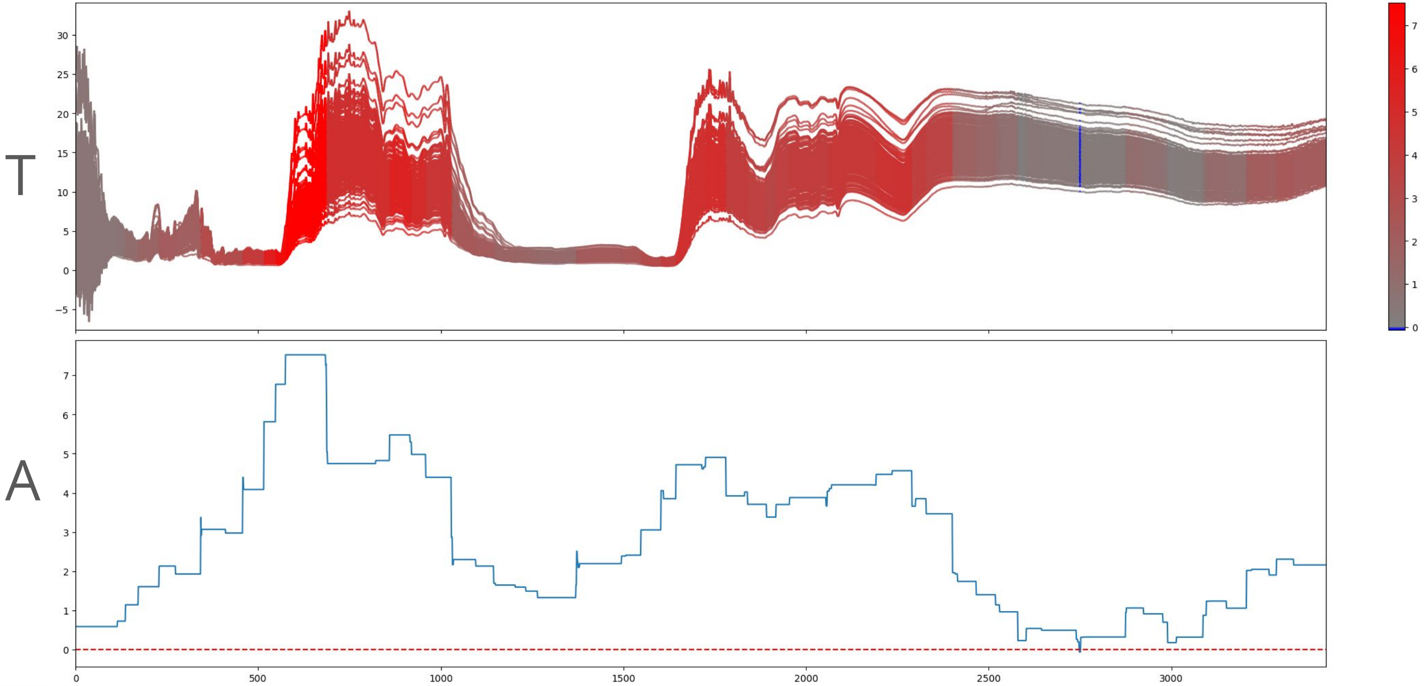
Results

Sample 09

True Lactose: 24



Results



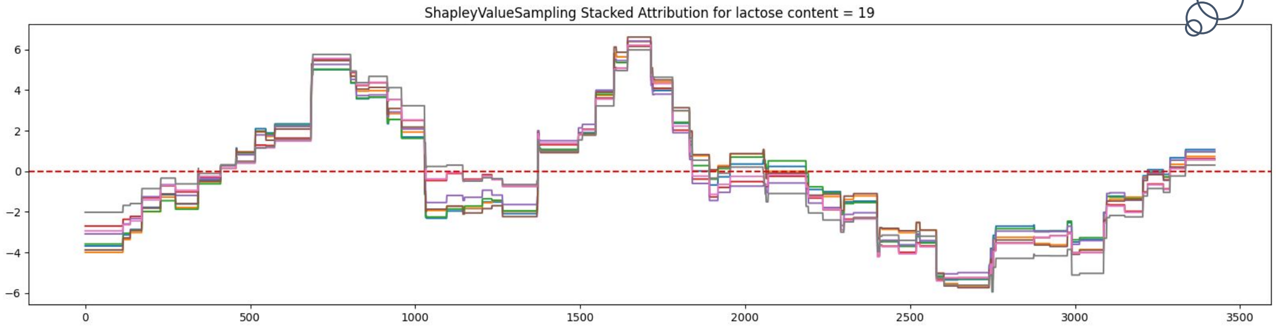
Sample 12

True Lactose: 72

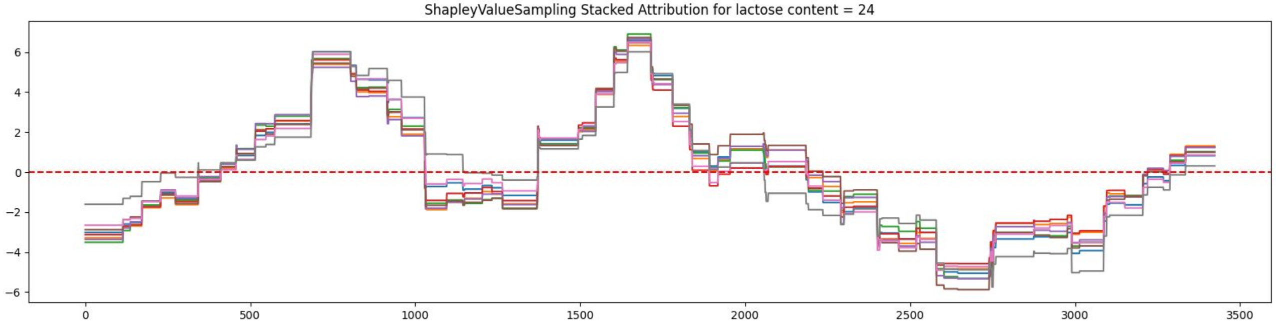
Insight

Explanation Profile

Let's stack the attributions of the samples with the same lactose content !



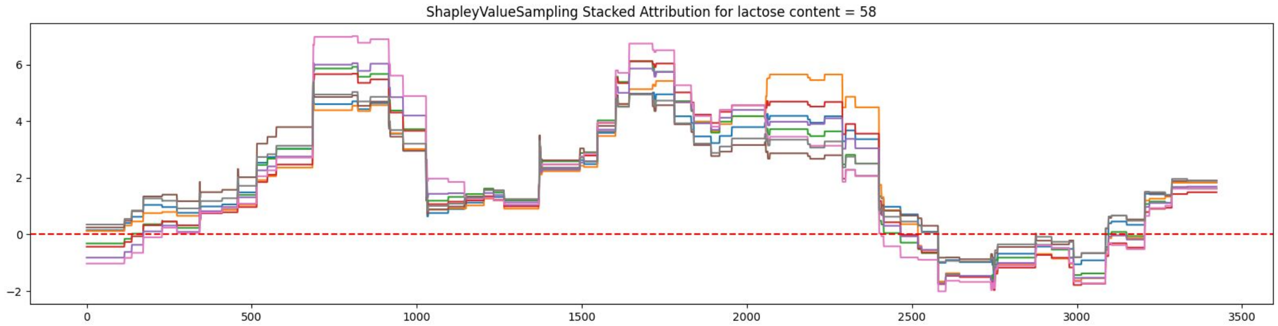
Lactose:19



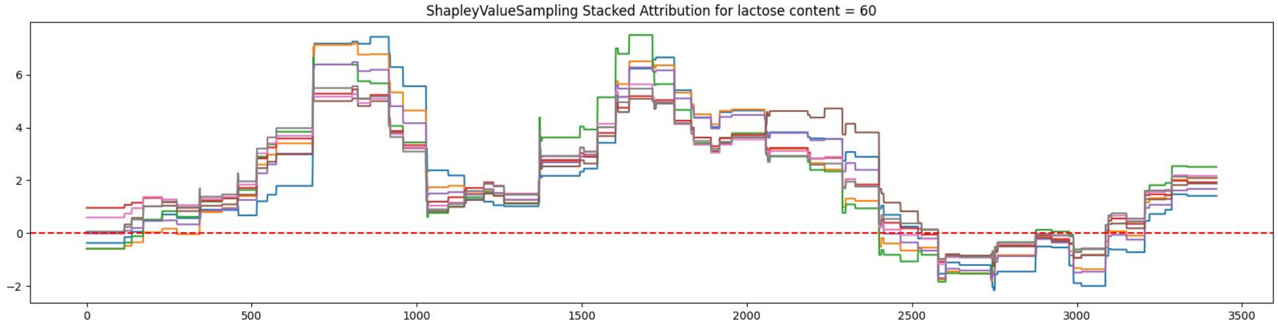
Lactose:24

Insight

Explanation Profile

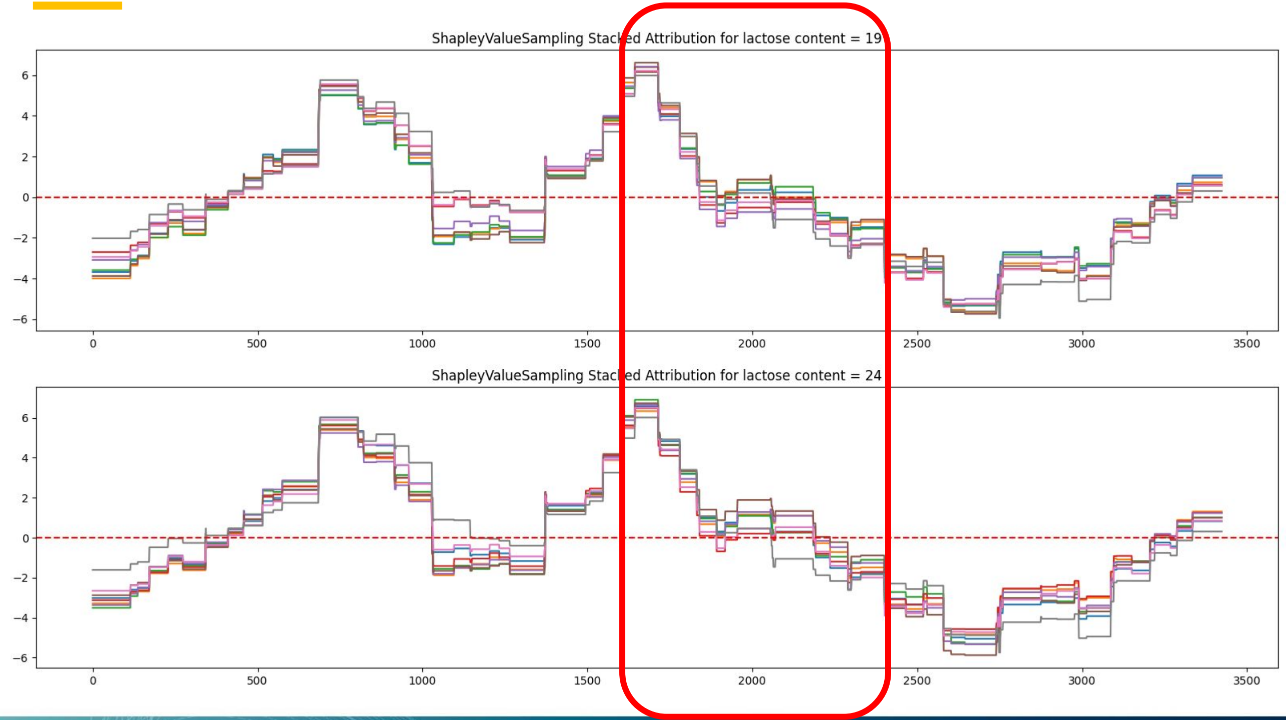


Lactose:58



Lactose:60

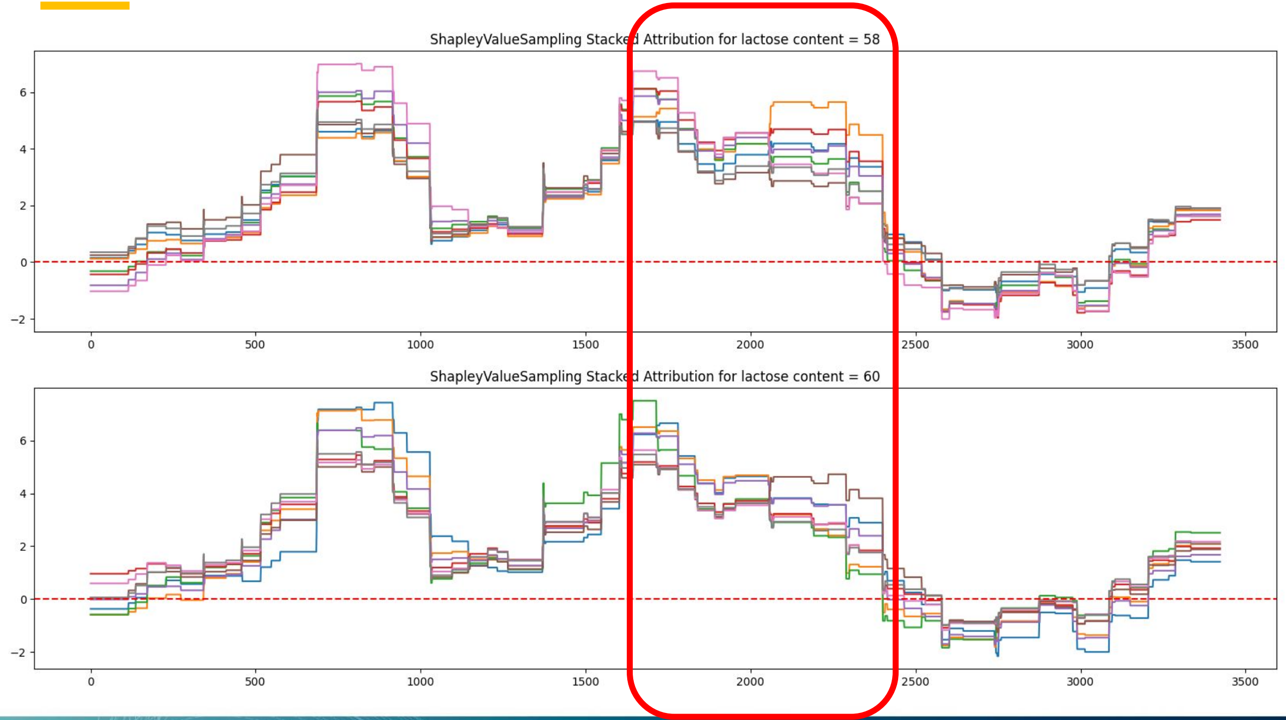
Explanation Profile



Lactose:19

Lactose:24

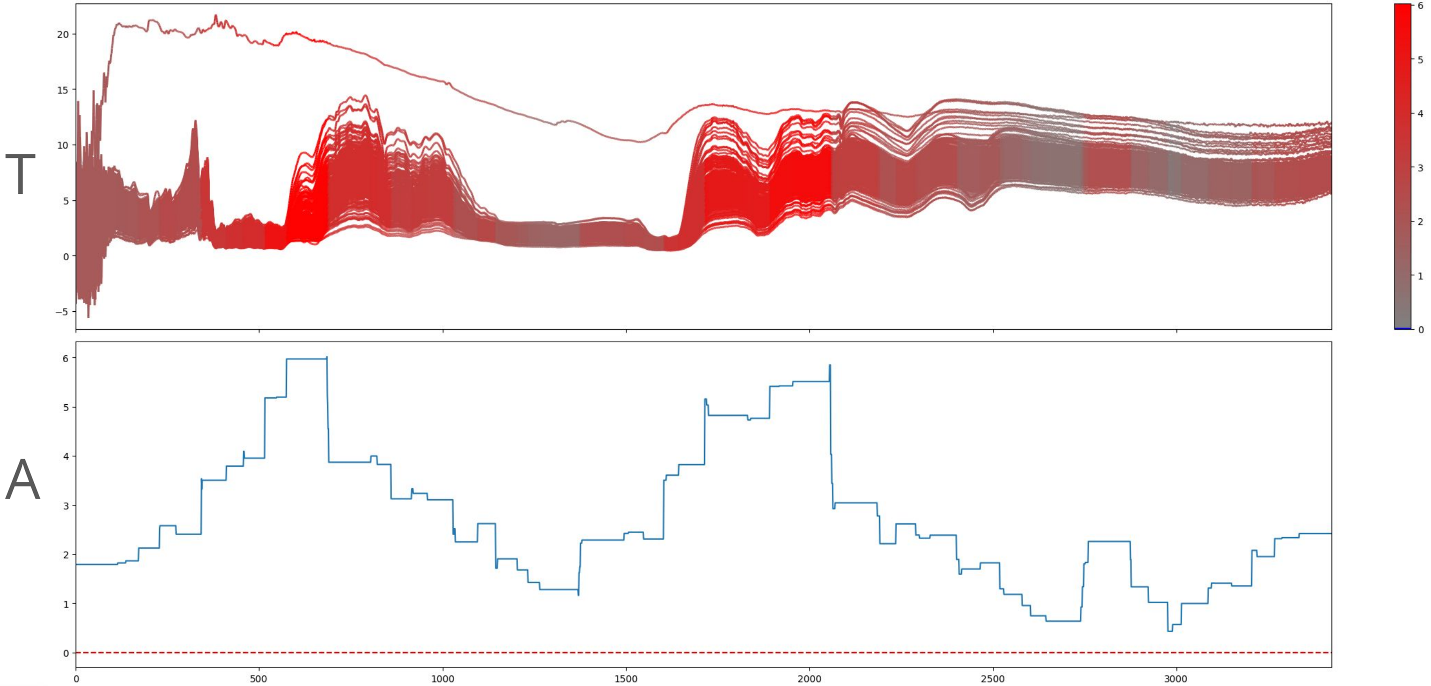
Explanation Profile



Lactose:58

Lactose:60

Test data



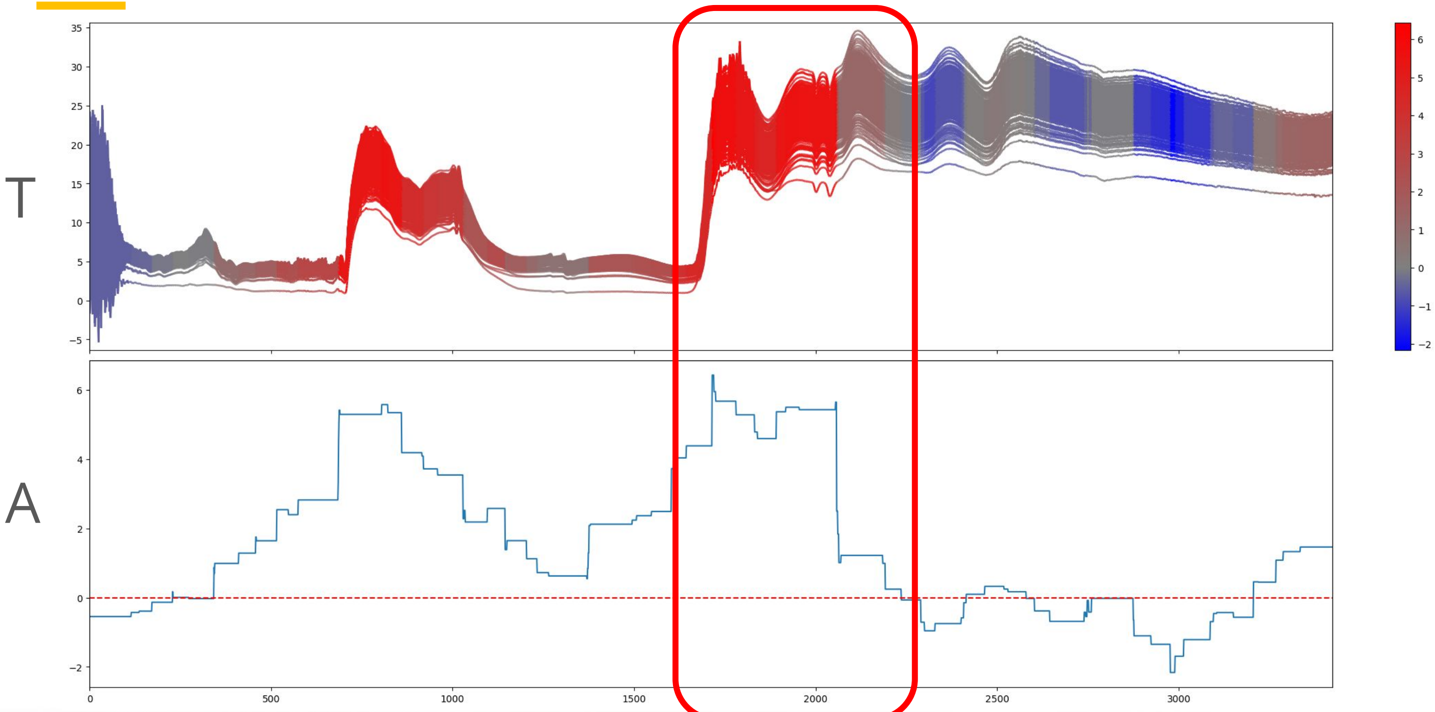
Sample 39

Human Prediction: 72

Model prediction: 72

True lactose: 72

Test data

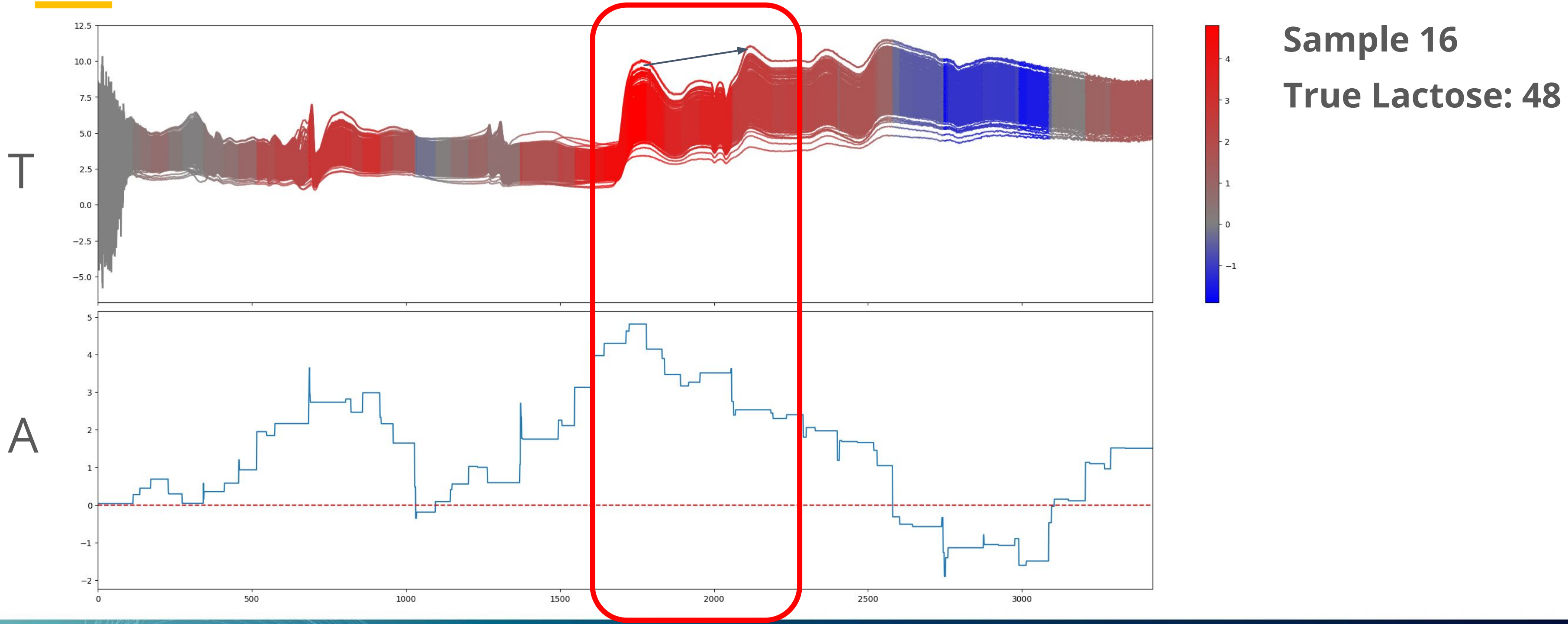


Sample 50

Human Prediction: 24

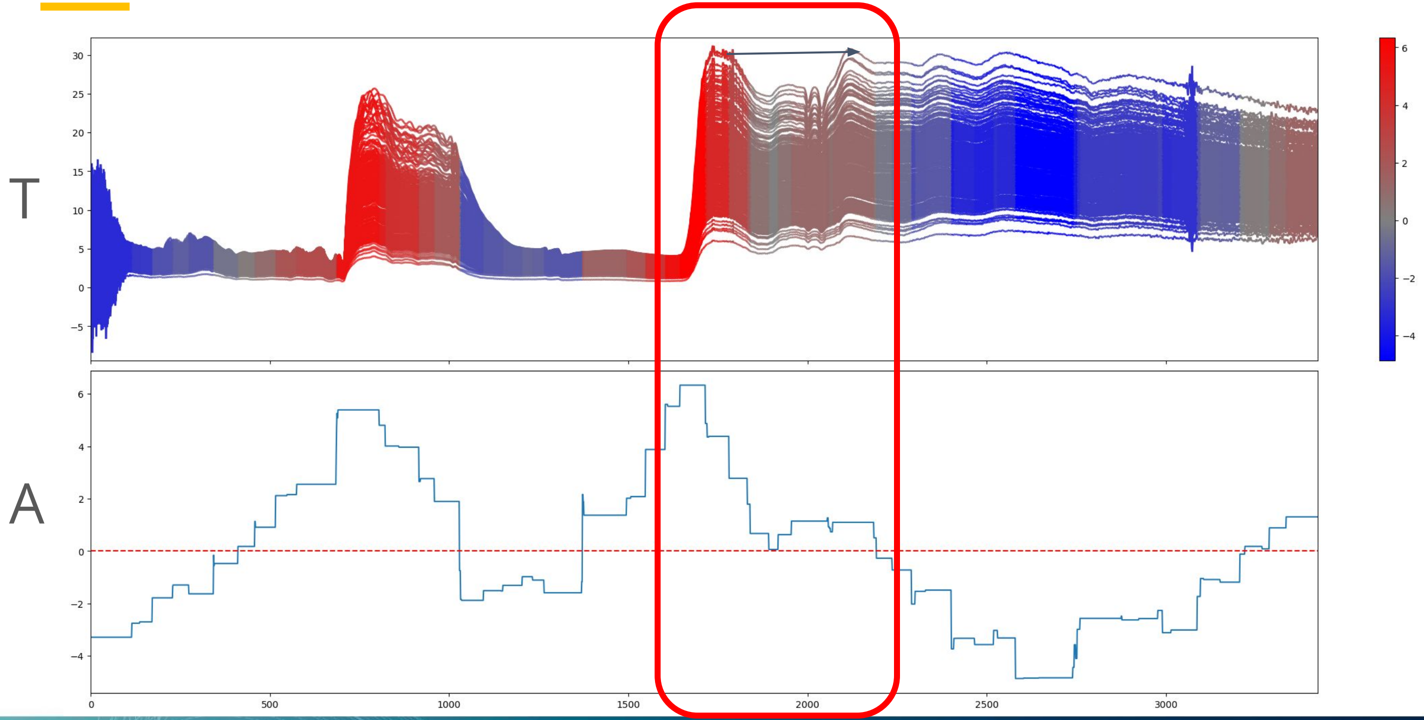
Model prediction: 53.14

Results



Sample 16
True Lactose: 48
T
A
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
5
4
3
2
1
0
−1
−2
0
500
1000
1500
2000
2500
3000

Results



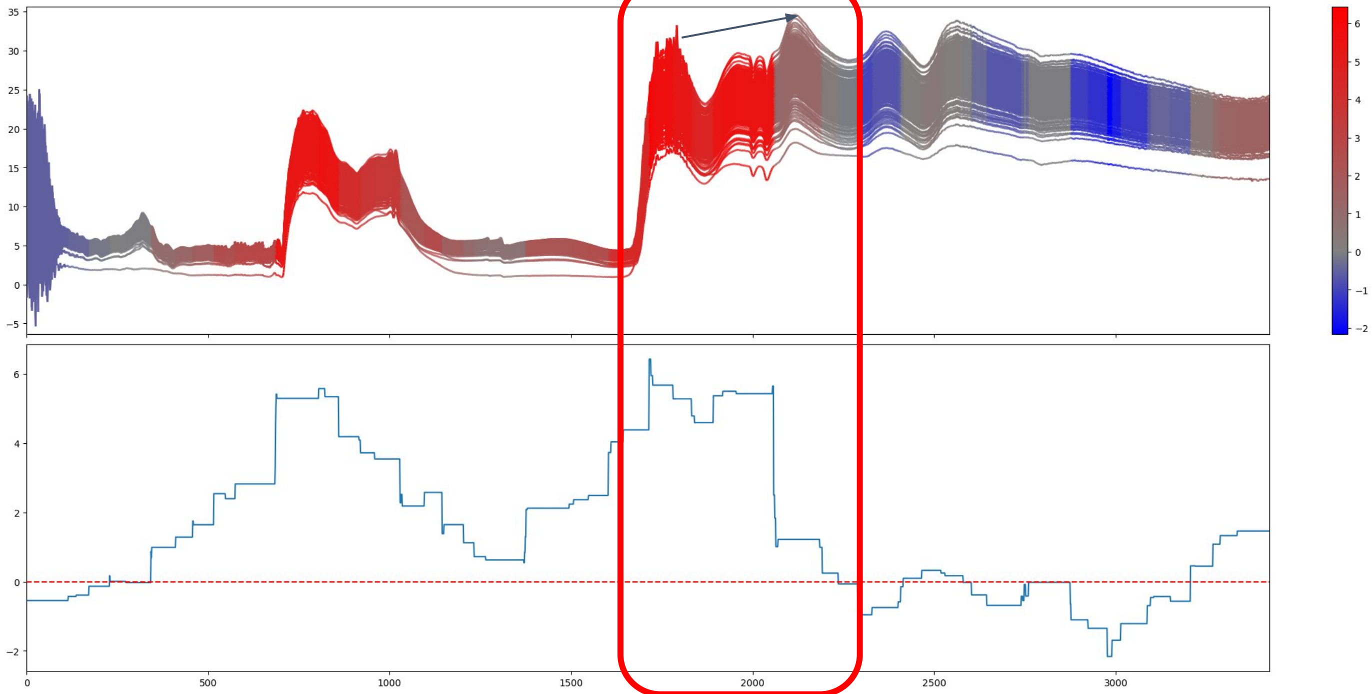
Sample 09

True Lactose: 24

Test data

T

A



Sample 50

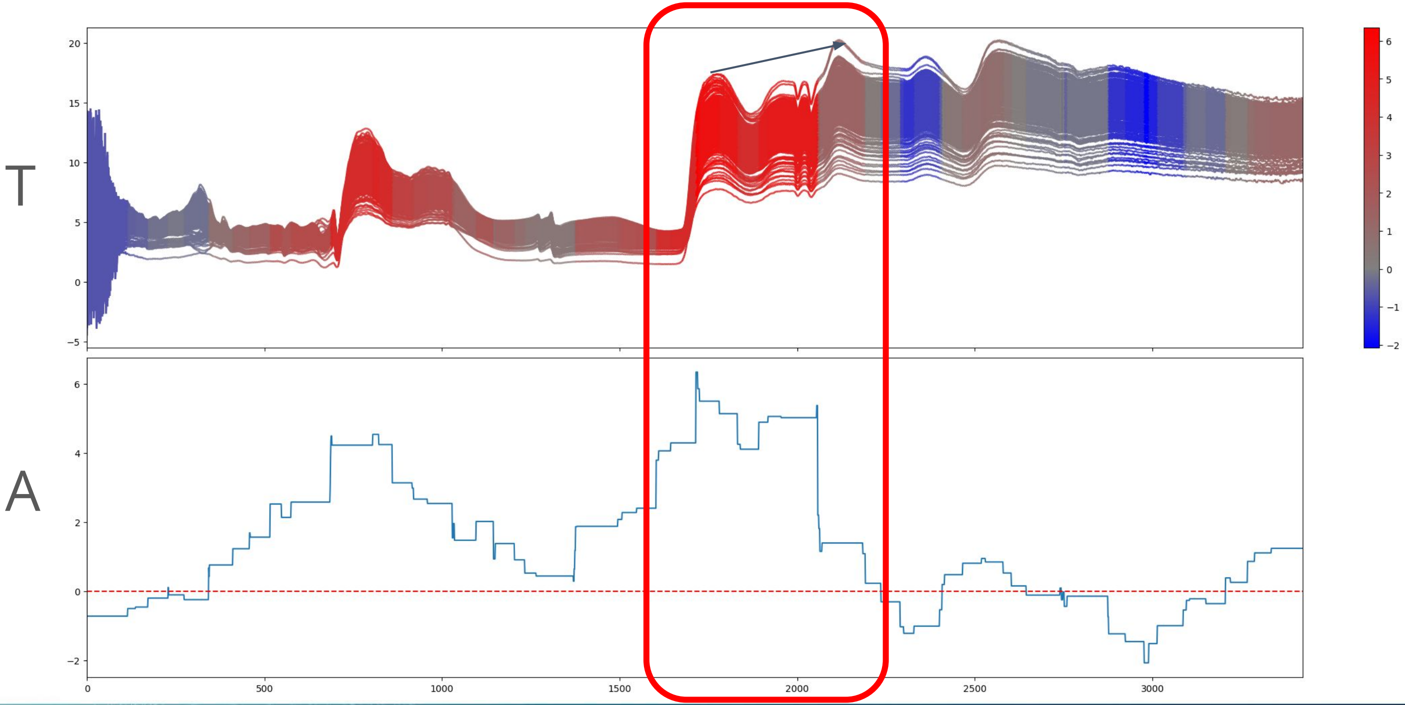
Human Prediction: 24

Model prediction: 53.14

Updated Human Prediction: 48

True lactose: 60

Test data



Sample 42

Human Prediction: 48

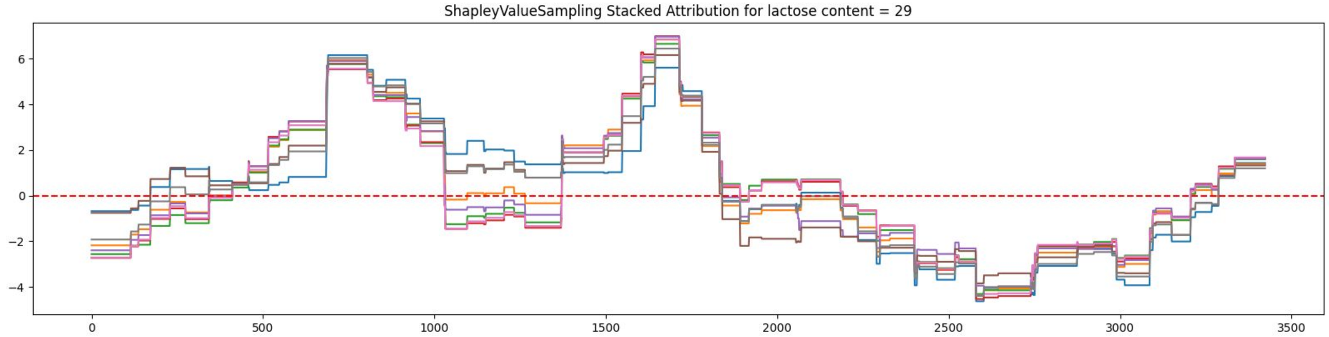
Model prediction: 49.24

True lactose: 24 !!!

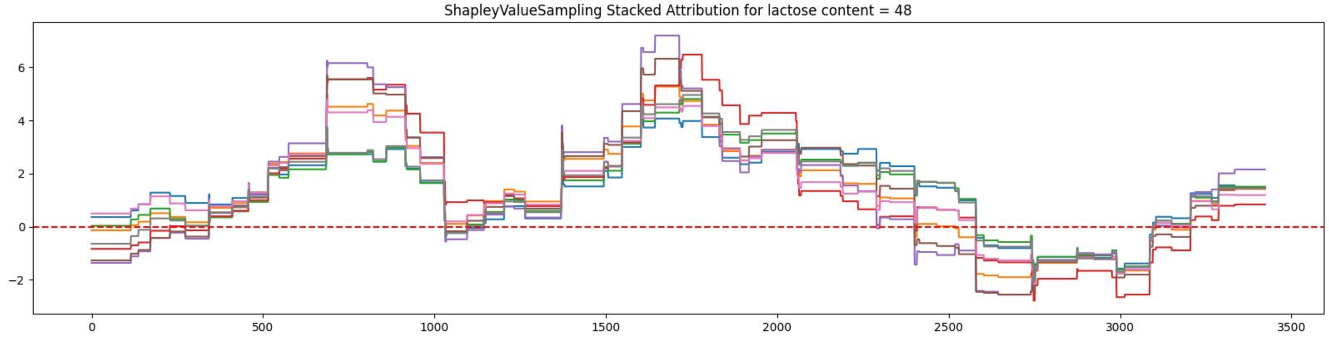
Take away

- We demonstrate how to use attribution methods to explain our models.
- The attribution can show us which parts of the input have impact on the model prediction.
- The attribution can be useful for understanding the model and the data.
- This is still a work in progress. We are curious if:
    - There is anything else we can learn from the attribution.
    - Is what we learn correct ? Is it valuable ? i.e. whether the domain experts find it accurate (sample 42 says otherwise) and whether they can use it for their works.
    - Can we use the attribution for any downstream task ? E.g. model enhancement, data labelling, data reduction etc.

Appendix



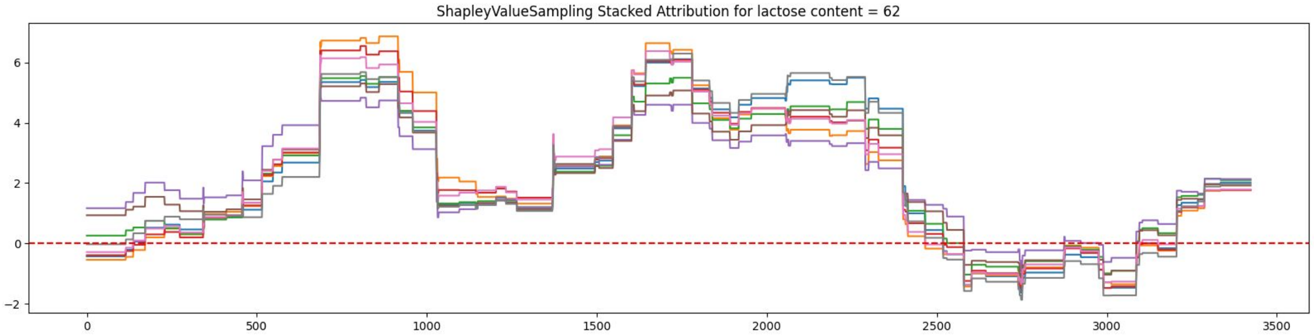
Lactose:29



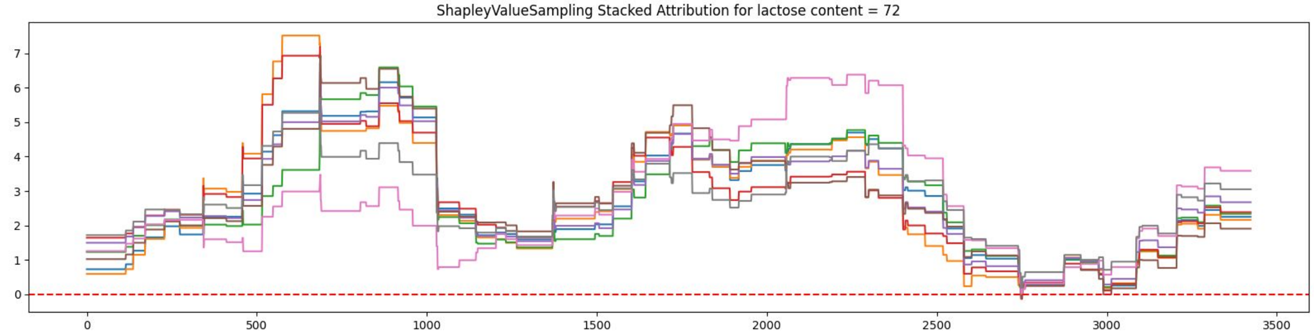
Lactose:58

Insight

Appendix



Lactose:62



Lactose:72