Time-Series Complexity into Understandable Prototypes: A Generic Approach to Machine Learning Explanations in Industrial Processes

mgr inż. Michał Kuk, PhD Candidate
prof. dr hab. inż. Grzegorz J. Nalepa
dr inż. Szymon Bobek

This presentation is funded from the XPM (Explainable Predictive Maintenance) project funded by the National Science Center, Poland under CHIST-ERA programme Grant Agreement No. 857925(NCNUMO – 2020/02/Y/ST6/00070)

Presentation plan

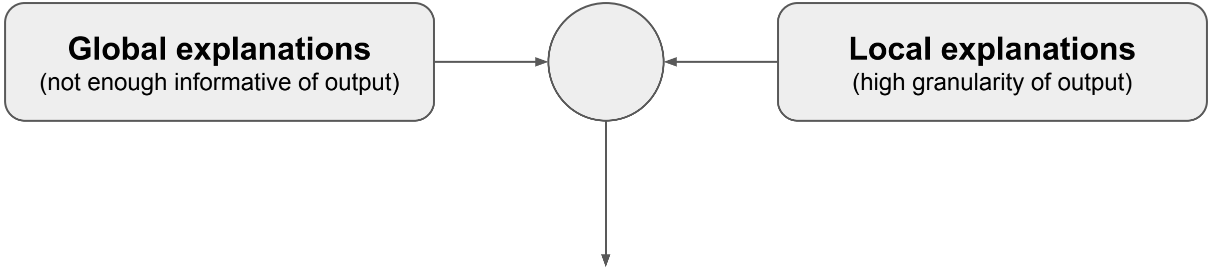
1. Feature importances as a tool for root cause analysis in time-series events
   a. need of explanations
   b. challenges with industrial assets
   c. anomaly identification and explanations as a kick-off to further analysis

2. Generic approach to Time-Series ML model explanations
   a. motivation and approach
   b. proposed solution
   c. preliminary study
   d. results presentation
   e. further steps

Feature importances as a tool for root cause analysis in time-series events

Need of explanations

Black-box: ML models are often seen as "black boxes", where their internal workings and decision-making processes are not transparent or understandable to human users. This opaqueness is a significant challenge in gaining trust and wider adoption of ML applications.



The "black-box" nature of ML models becomes more complex with time-series data due to its dynamic characteristics. Local explanations, tied to specific time points, are often hard to interpret. To tackle this, we propose summarizing these explanations into **understandable "prototypes",** effectively making the complex decision-making process more transparent and actionable for users.

Feature importances as a tool for root cause analysis in time-series events

Challenges with industrial assets

1. Failures in industrial assets are usually rare events, occurring after extended periods of seamless operation. This scarcity of failure instances presents a unique challenge for machine learning models

2. Given the rarity of these failures, ML models often focus on anomaly detection to predict possible breakdowns.

3. When a potential anomaly is flagged, explanations are needed to validate and understand these rare predictions. This helps operators trust the model's predictions and take targeted preventive actions

Feature importances as a tool for root cause analysis in time-series events

Anomaly identification and explanations

What we have done?

Dataset: Steel coil production process

Target: Unsupervised learning

Unsupervised learning

↓

Autoencoder — Anomaly detection based on reconstruction error

↓

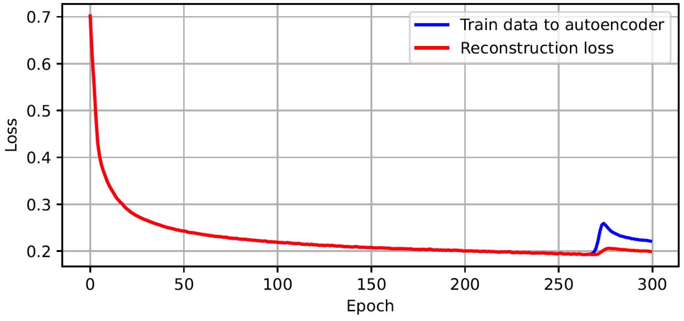
Classification task — XGBoost algorithm (building a model needed for explanation algorithm)

↓

Explanations

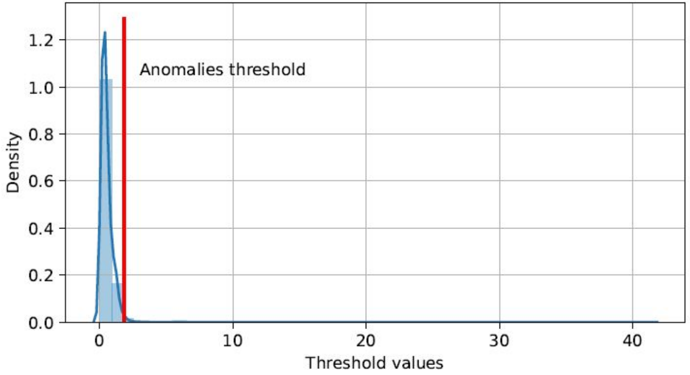
Feature importances as a tool for root cause analysis in time-series events

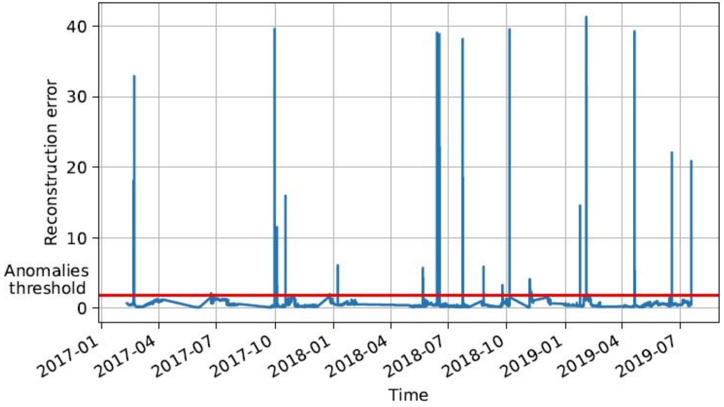
Anomaly identification and explanations



Dataset details:
Dataset shape: 35 features, 24 000 instances

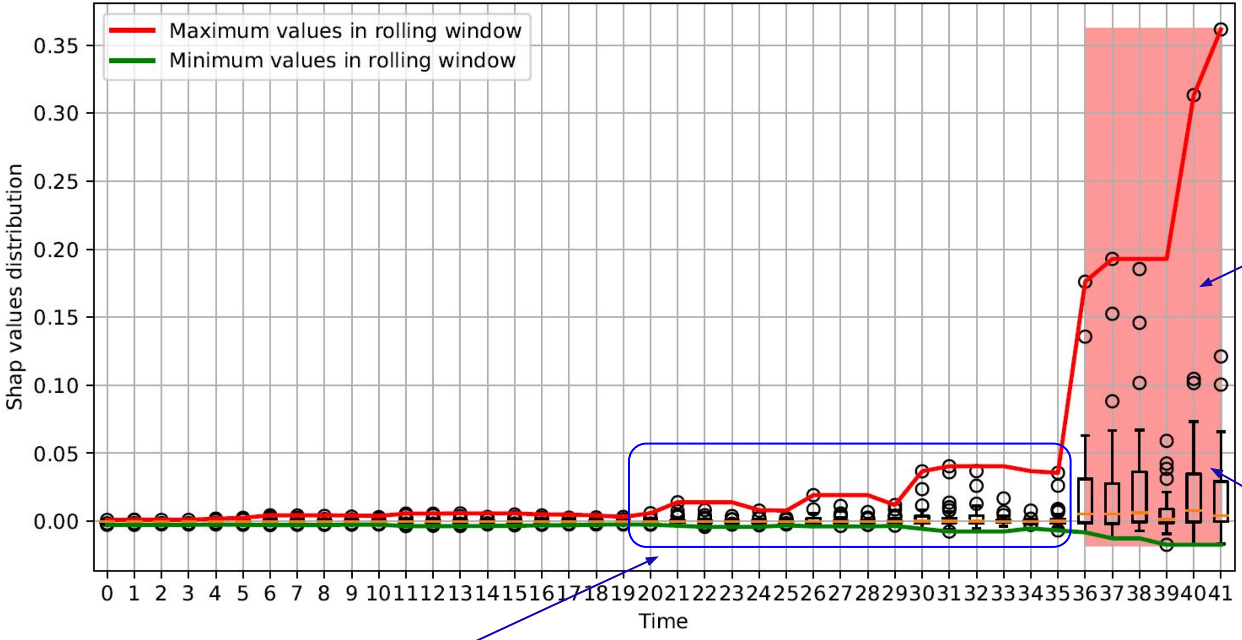
Autoencoder details:
Type of autoencoder: based on convolutional layers
Number of layers: 6
Latent space shape: 4
Activation function: ELU
Reconstruction error threshold: 0.99 quantile of RE





Feature importances as a tool for root cause analysis in time-series events

Anomaly identification and explanations



Marked anomaly

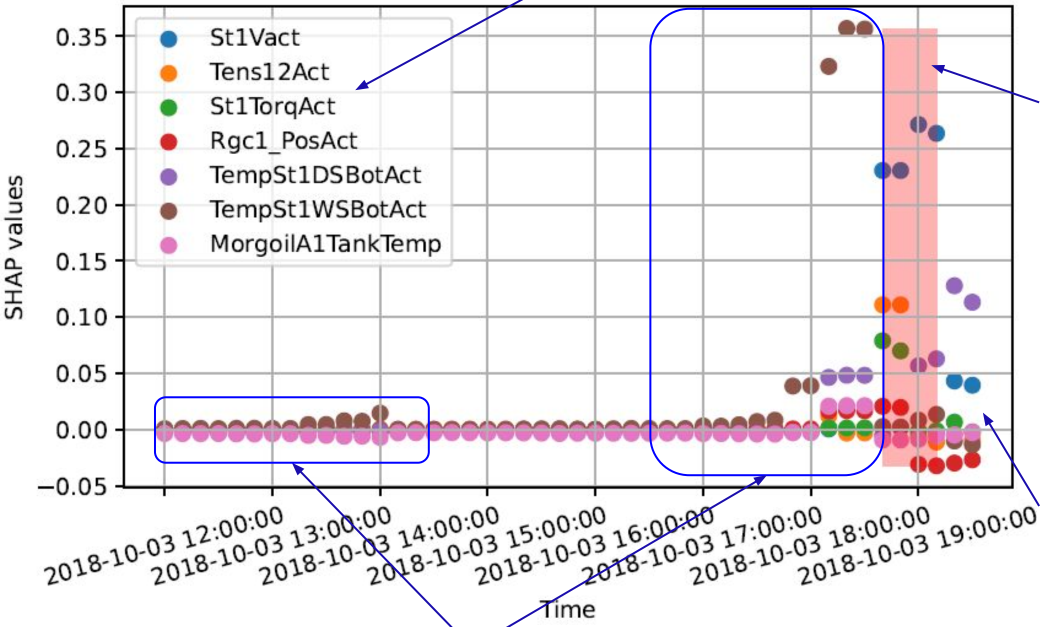
(Data points) - SHAP values from all features

Something happens (increase in SHAP values)

Feature importances as a tool for root cause analysis in time-series events

Anomaly identification and explanations

Features which contribute



Marked anomaly

(Data points) - SHAP values from important features

Something happens (increase in SHAP values)

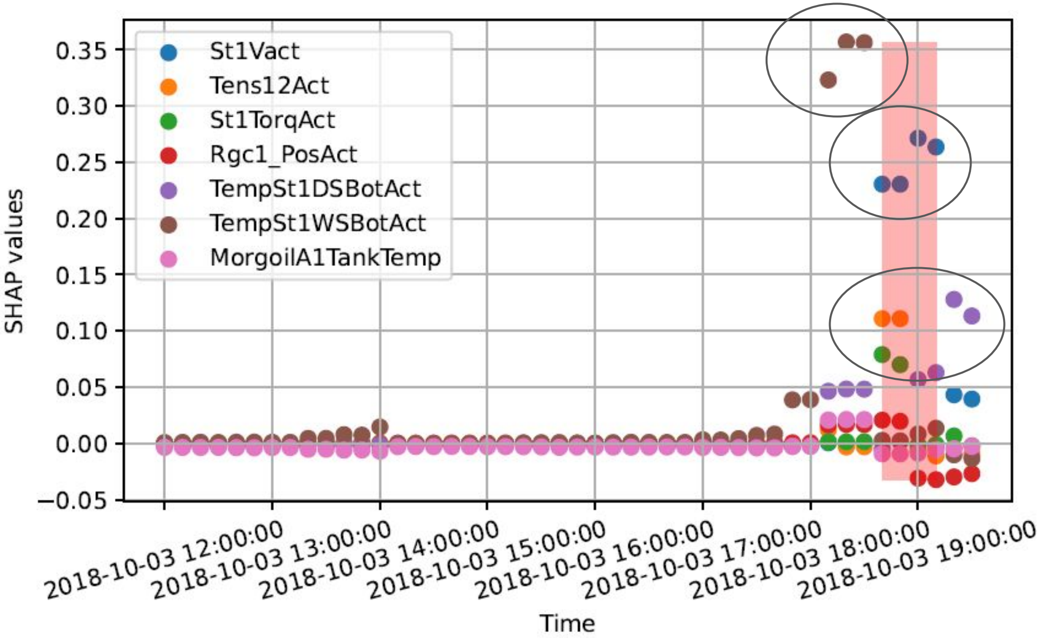
SHAP explanations remark:
Values higher than 0 - feature force model to predict positive class (failure/anomaly)
Values lower than 0 - feature force model to predict negative class (normal work)

Identifying of the most important features:

1. Analysis of the SHAP values in rolling windows
2. Calculating the distribution of features contribution
3. 0.8 quantile cut-off value

Feature importances as a tool for root cause analysis in time-series events

Further steps (such explanations problems)



Taking into account the stability of such explanations it is not clear:

1. Which feature contribute the most
2. In which samples specific features contribute in prediction and how much
3. What with the feature which indicates early symptoms and later the importance is relatively low

Difficulties in understanding (solution?)

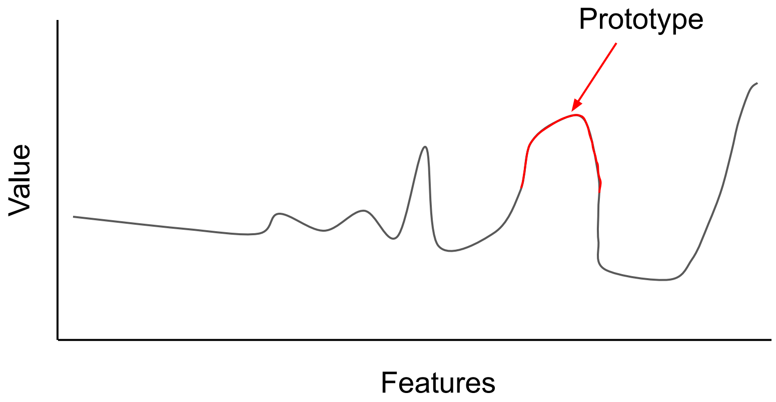
Generic approach to ML model explanations

Motivation and approach

Motivation:
Taking into account the stability of such explanations it is not clear:

1. Which feature contribute the most
2. In which samples specific features contribute in prediction
3. What with the feature which indicates early symptoms
4. Many others…

Approach:
Generate summary of the data in the form of prototypes



We are looking for answer:
If such prototype occurs than you probably have an failure indication

Remark: On the chart is presented only one instance

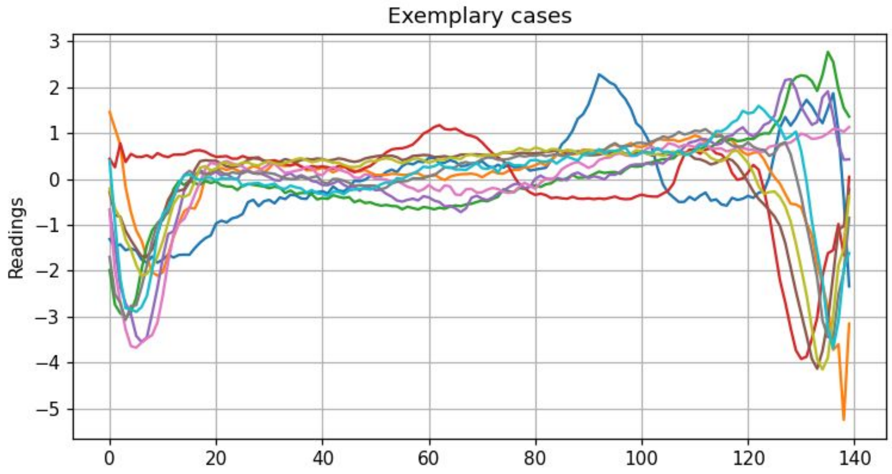
Generic approach to ML model explanations

Proposed solution

Step by step method:

1. Building a classifier
2. Calculating SHAP values
3. Detecting change points on SHAP values
4. Clustering using DTW (dynamic time warping) metric
5. Converting task to prototype manner
6. Building a classifier on prototypes
7. Identification of prototypes

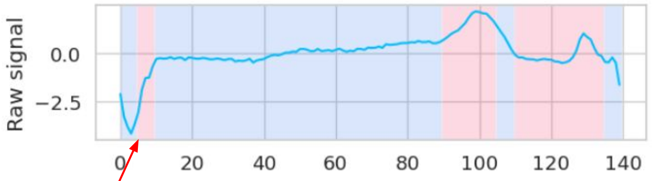
Dataset:
The ECG dataset is composed of two collections of heartbeat signals derived from two famous datasets in heartbeat classification



Generic approach to ML model explanations

Preliminary study

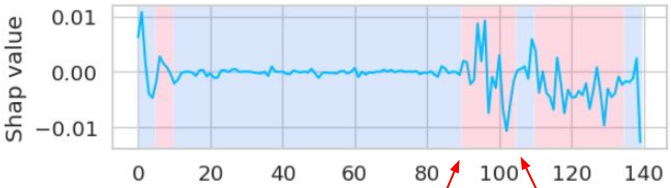
1. From 5 target labels we simplified the task to two categories - sick or health
2. We builded a classifier
3. Based on the classification model we generated an SHAP values



Indicated shifts based on shap value

Change point detection - Reptures package

Parameters:
search method: Pelt
model: rbf
penalty: 1.2



Begin of segment

End of segment

The main goal is to split the signal represents as SHAP values for a chunk of the data.

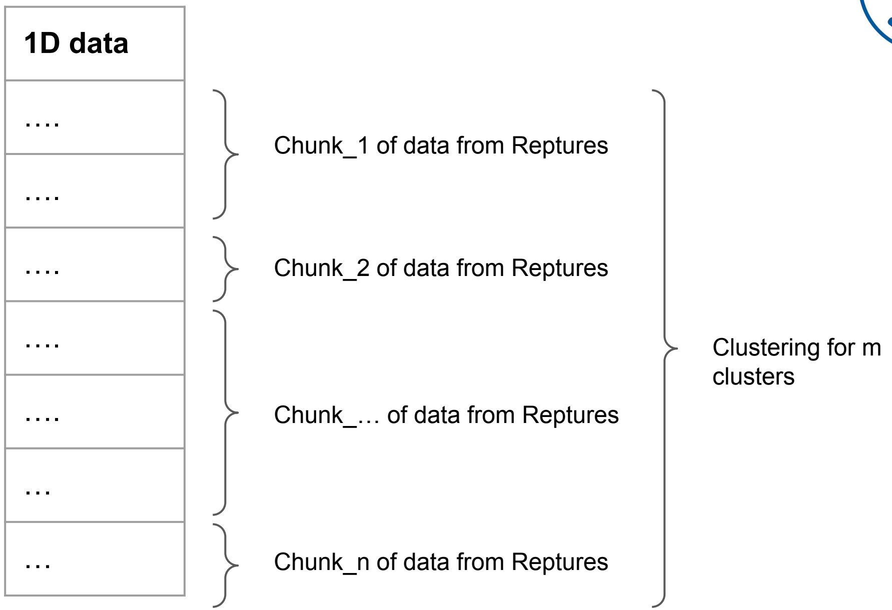
Detecting change points on SHAP values | Clustering using DTW (dynamic time warping) metric | Converting task to prototype manner | Building a classifier on prototypes | Identification of prototypes

Generic approach to ML model explanations

Preliminary study

Clustering using DTW (dynamic time warping) metric - TimeSeriesKMeans

Parameters:
number of clusters: [2-20]
metric: DTW
cluster separation metric: silhouette score



Detecting change points on SHAP values

Clustering using DTW (dynamic time warping) metric

Converting task to prototype manner

Building a classifier on prototypes

Identification of prototypes

Generic approach to ML model explanations

Preliminary study

Converting task to prototype manner

	Segment_1		Segment_2			Segment_3	Segment_…		Segment_n
	Feature_1	**Feature_2**	**Feature_3**	**Feature_4**	**Feature_5**	**Feature_6**	**Feature_7**	**Feature_8**	**Feature_9**
Instance_1									
Instance_2									

SHAP values data

Start End Start End ……………………………….

Result of the this step:
Instance 1 - is divided in to chunks:
[Feature_1, Feature_3, Feature_6, Feature_7, Feature_9] (only starts of the each segment)

Instance 2 - can be divided into another chunks

Detecting change points on SHAP values | Clustering using DTW (dynamic time warping) metric | Converting task to prototype manner | Building a classifier on prototypes | Identification of prototypes

Generic approach to ML model explanations

Preliminary study

Converting task to prototype manner

	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9
Instance_1									

SHAP values data

Use clustering model on raw signal data to predict if the chunk of raw signal data belongs to specific cluster or not

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6
Instance_1	0	1	1	0	1	1
Instance_2	1	0	1	1	0	1

Results of clustering (less dimension)

Raw signal data

Detecting change points on SHAP values

Clustering using DTW (dynamic time warping) metric

Converting task to prototype manner

Building a classifier on prototypes

Identification of prototypes

Generic approach to ML model explanations

Preliminary study

Converting task to prototype manner

Use clustering model on raw signal data to predict if the chunk of raw signal data belongs to specific cluster or not

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6
Instance_1	0	1	1	0	1	1
Instance_2	1	0	1	1	0	1

Results of clustering (less dimension)

Raw signal data

E.g. this prototype does not exist in this instance

E.g. this prototype exists in this instance

Detecting change points on SHAP values

Clustering using DTW (dynamic time warping) metric

Converting task to prototype manner

Building a classifier on prototypes

Identification of prototypes

Generic approach to ML model explanations

Preliminary study

Building a classifier on prototypes

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6
Instance_1	0	1	1	0	1	1
Instance_2	1	0	1	1	0	1
Instance_3	0	1	1	0	0	0

Label
0 - health
1 - sick
1 - sick

XGBoost classifier

Parameters:
max depth: [0-9]

Results:
Accuracy: 0.89

Marge this tables and learn classifier to
feed explainer algorithm

Detecting change points on SHAP values

Clustering using DTW (dynamic time warping) metric

Converting task to prototype manner

Building a classifier on prototypes

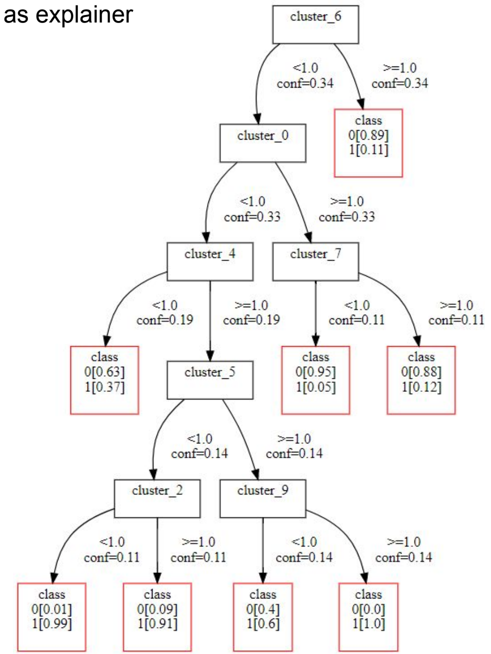
Identification of prototypes

Generic approach to ML model explanations

Preliminary study

Identification of prototypes

Lux algorithm as explainer



Examples of obtained rules:

IF cluster_1 >= 1.0 AND cluster_0 >= 1.0 THEN class = 1
IF cluster_7 >= 1.0 AND cluster_4 < 1.0 AND cluster_6 < 1.0 AND cluster_0 < 1.0 THEN class = 0
IF cluster_4 >= 1.0 AND cluster_6 >= 1.0 THEN class = 0

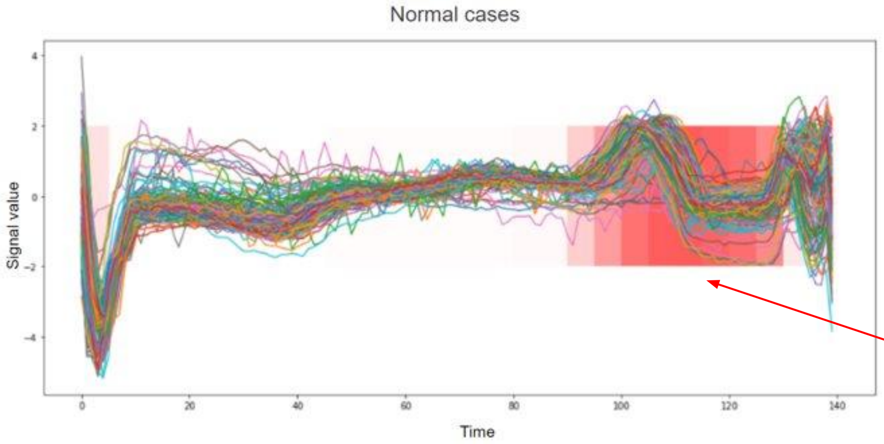
For each of instance the set of rules consist of prototypes has been generated

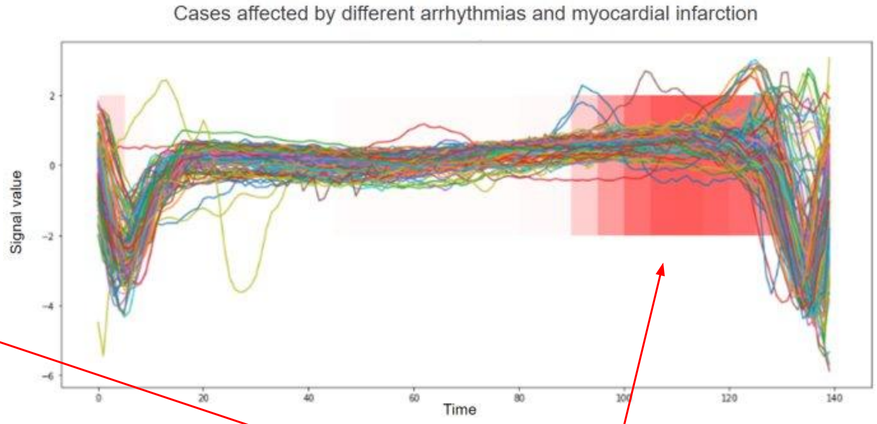
Detecting change points on SHAP values | Clustering using DTW (dynamic time warping) metric | Converting task to prototype manner | Building a classifier on prototypes | Identification of prototypes

Generic approach to ML model explanations

Results presentation

Analysis of rules obtained for each of instance





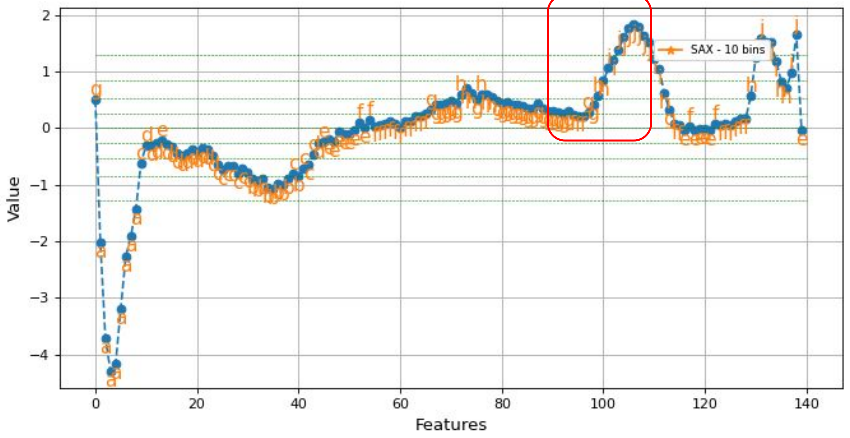
Marked segments which were indicated by the generated rules (LUX algorithm) presented on all analyzed cases.

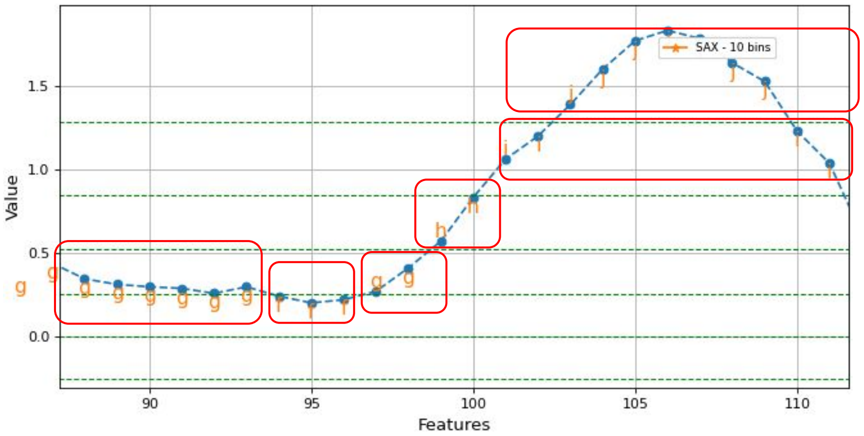
The explainable algorithm found the segment which the most differentiates the normal and sick cases. The indicated segments could be treated as a prototype which in a human understanding way presents why the algorithm classifies the signal for normal or not normal ECG.

Generic approach to ML model explanations

Future work

1. **Symbolic Aggregate approXimation (SAX)**





2. ProtoPNet
3. Evaluation on real industrial dataset

Thank You for Your attention