Globally-Consistent Rule-Based Summary-Explanations for Machine Learning Models: Application to Credit-Risk Evaluation[1]

MI2 DataLab Summer Research Seminar 2023

Mikołaj Spytek supported by Mateusz Krzyziński

[1]Cynthia Rudin, Yaron Shaposhnik

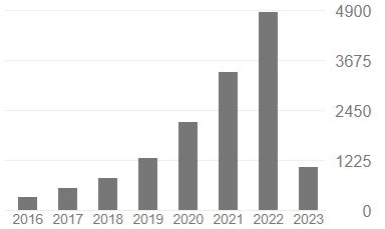
Cynthia Rudin

PI, Interpretable Machine Learning Lab, Duke University

Rudin, C. ***Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead***. *Nature Machine Intelligence* 1, 206–215 (2019).

Fisher, A., Rudin, C., Dominici F. ***All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously***. Journal of Machine Learning Research 20.177 (2019): 1-81.

	Wszystkie	Od 2018
Cytowania	15885	13553
h-indeks	51	42
i10-indeks	112	89



What's wrong with local explanations?

What's wrong with local explanations?

- All frequently used explanations do not have the global-consistency property.
- In cases where explanations are required by regulations, we're not sure that they agree with the underlying model.

Quote from the paper:
"The most serious mistake in applying explanations is arguably that **explanations are generally not consistent with the underlying model** they are trying to explain. For instance, imagine a person being denied credit by a model, receiving an explanation such as "credit history not greater than 10 years." However, a different person could have a credit history less than 10 years and yet could be granted credit. This is a case where the explanation is not globally-consistent with the underlying model. In some cases, the explanation could actually produce **the opposite prediction** as the global model, which means it is not trustworthy."

Proposed solution

- A method for calculating **globally-consistent** *summary-explanations.*
- They apply to only a local part of the search space, however they agree with the underlying model for all observations in this subspace
- The explanations are **rule-based**, that is they restrict the subspace by a set of rules based on the predictors.

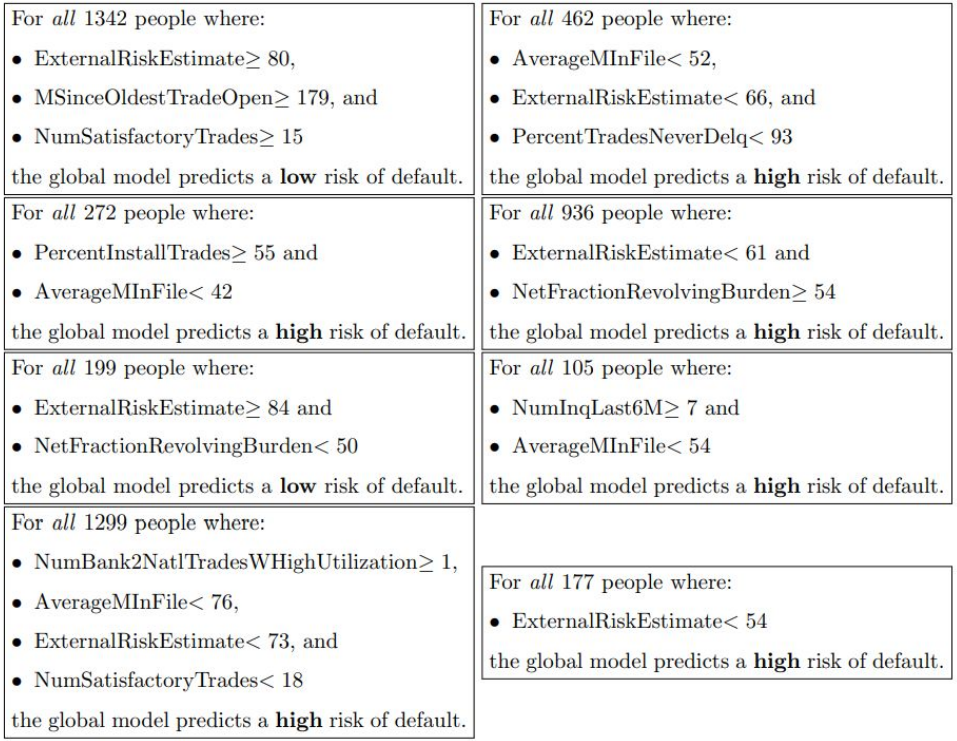
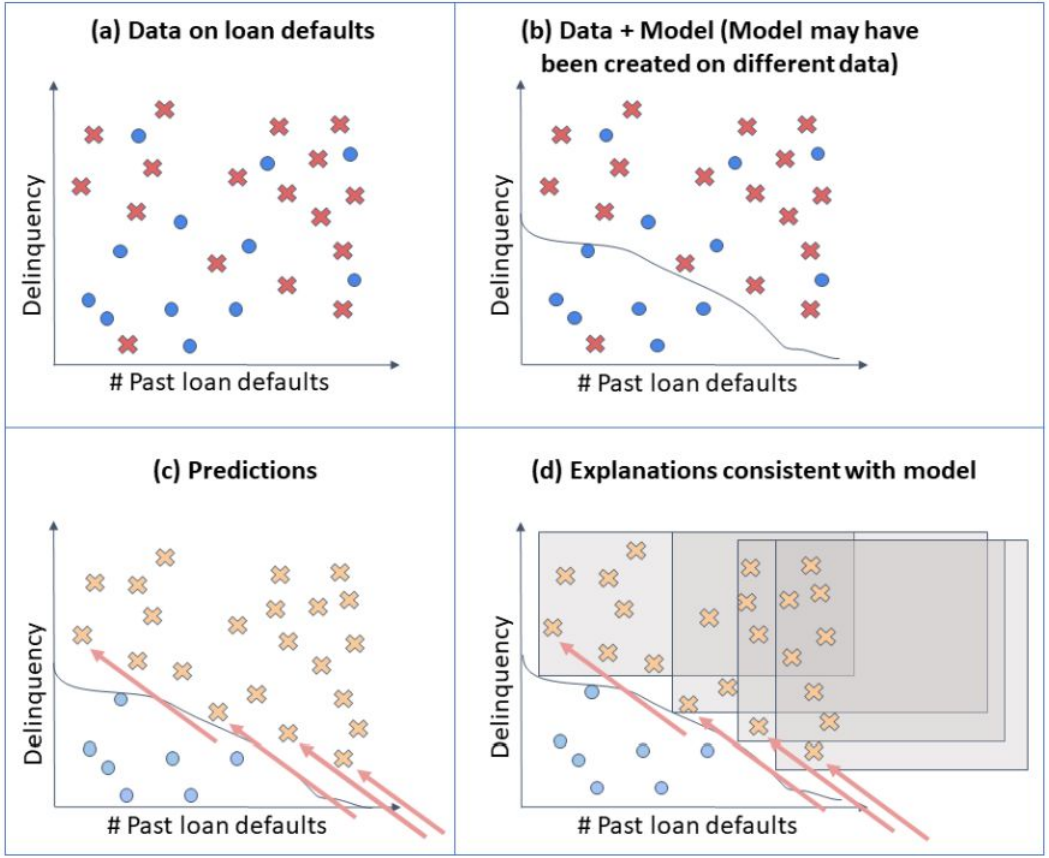


Figure 1 from Rudin, C., Shaposhnik, Y. "Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation."

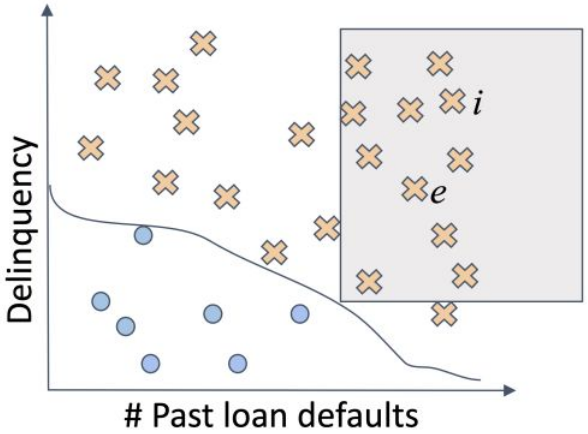

(a) Data on loan defaults
Delinquency
Past loan defaults
(b) Data + Model (Model may have been created on different data)
Delinquency
Past loan defaults
(c) Predictions
Delinquency
Past loan defaults
(d) Explanations consistent with model
Delinquency
Past loan defaults

Globally-consistent local summary-explanations

★ **relevance**
the summary explanation for the point x_e must agree with the model's prediction for x_e

★ **global consistency**
if the explanation of point x_e covers the point x_i, then the point x_i must have the same prediction of the model as x_e

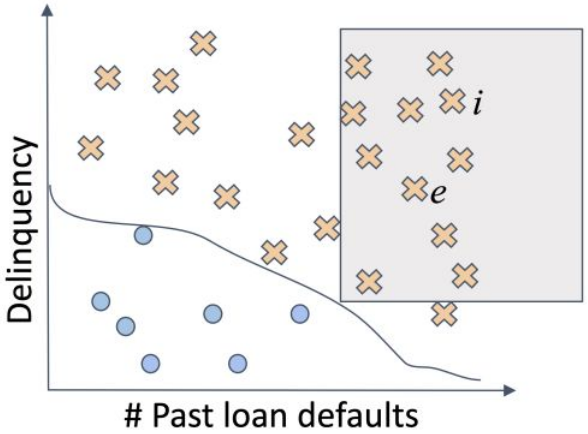
★ **interpretability**
the explanation is interpretable, that is we know what subspace is covered by it



How to assess the quality of an explanation?

complexity
a measure of (the inverse of) interpretability, can be different for different methods.
In this paper: number of rules |R|

support
the number of observations covered by the explanation, greater values increase trust and confidence in the explanation



Optimization problem formulation

OptConsistentRule

$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
&\qquad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && \text{Then } \exists p \in R : x_{i,p} < \tau_p && \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

Optimization problem formulation

OptConsistentRule

number of observations covered by epxlanation

$$
\begin{aligned}
\max_{R,\tau} \quad & w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
\text{s.t.} \quad & \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
\forall i \in N : \quad & \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& \text{Then } \exists p \in R : x_{i,p} < \tau_p \\
& |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

Optimization problem formulation

OptConsistentRule

number of rules

$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
&\quad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && \text{Then } \exists p \in R : x_{i,p} < \tau_p && \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

Optimization problem formulation

OptConsistentRule

weights for importance of support and complexity

$$
\begin{aligned}
\max_{R,\tau} \quad & w_s \; \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
\text{s.t.} \quad & \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
\forall i \in N : \quad & \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& \text{Then } \exists p \in R : x_{i,p} < \tau_p \\
& |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

Optimization problem formulation

OptConsistentRule

$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{threshold for predictor p} \\
&\qquad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && \text{Then } \exists p \in R : x_{i,p} < \tau_p && \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

Optimization problem formulation

OptConsistentRule

$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
&\quad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

prediction for observation *i*

Optimization problem formulation

OptConsistentRule

$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
&\qquad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

prediction for observation *e*

Optimization problem formulation

OptConsistentRule

$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
&\qquad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && \text{Then } \exists p \in R : x_{i,p} < \tau_p && \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

Optimization problem formulation

OptConsistentRule

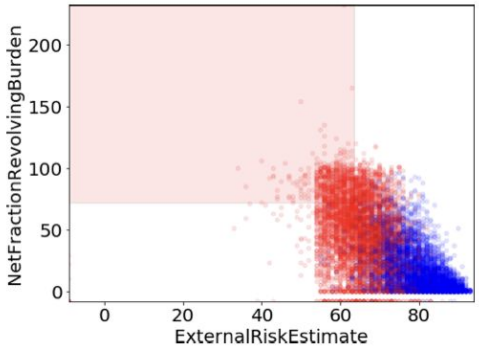
$$
\begin{aligned}
&\max_{R,\tau} && w_s \cdot \Gamma_s(h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}) - w_c \cdot |R| && \text{(maximize support, minimize complexity)} \\
&\text{s.t.} && \forall p \in R : x_{e,p} \geq \tau_p && \text{(summary-explanation is relevant)} \\
&\qquad \forall i \in N : && \text{If } h^{\text{global}}(\mathbf{x}_i) = 1 - h^{\text{global}}(\mathbf{x}_e) && \text{(summary-explanation is consistent)} \\
& && \text{Then } \exists p \in R : x_{i,p} < \tau_p && \\
& && |R| \leq p_c && \text{(interpretability)}
\end{aligned}
\tag{3}
$$

maximal number of rules

Geometric interpretation

Globally-consistent rule:

- is a hyperbox in the feature space
- contains the observation we wish to explain
  and only other observations that are similarly labeled



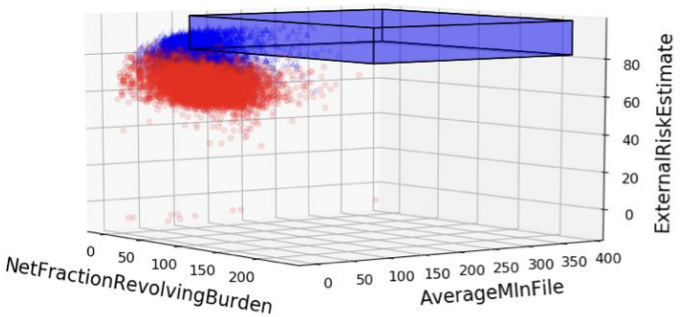


Figure 4: Graphical illustration of two rules (e.g., the left rule is $h_e(\mathbf{x}) = \mathbb{1}[\text{ExternalRiskEstimate} \leq 63 \ \& \ 73 \leq \text{NetFractionRevolvingBurden}]$). These rules are summary-explanations of all the points they contain, and are based on features that are interpretable within their context.

Considerations of computational complexity

OptConsistentRule generalizes
the Minimum Set Cover problem
(proved in paper)

+

Minimum Set Cover problem is NP-hard
(proved before)

→ OptConsistentRule is NP-hard

Considerations of computational complexity

OptConsistentRule generalizes
the Minimum Set Cover problem
(proved in paper)

+

Minimum Set Cover problem is NP-hard
(proved before)

→ OptConsistentRule is NP-hard

Quote from the paper:
While MinSetCover is a theoretically difficult problem, from a practical standpoint, it **can be easily implemented and solved** using current computing technologies and Integer Programming (IP) solution techniques.

BinMinSetCover and BinMaxSupport

BinMinSetCover is an algorithm to find the explanation, which covers the explained observation with the **minimal number of rules** in a **binary dataset.** It is solved with Integer Programming.

Then BinMaxSupport algorithm is described, where the minimality constraint is relaxed, **trading the number of rules for increased support**.

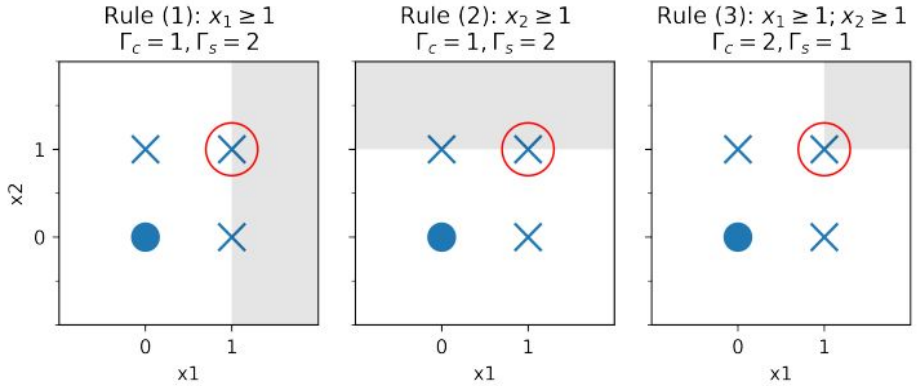


Figure 6: A simple example of rule-based summary-explanations for a binary dataset. The red circle denotes the observation being explained (which in this case, is also part of the data), while the shaded area depicts the part of the feature space where the summary-explanation applies. Rule (1) and Rule (2) both dominate Rule (3) because they cover more than just the one point of Rule (3) while using fewer thresholds.

What about continuous datasets?

The explanations are useful only if we can apply them to a **general dataset** without many assumptions. Manual binarization of all variables is tedious and leads to model performance loss.

For continuous datasets the authors propose the following procedure:

1. **Reduce** the continuous dataset **to a binary dataset**.
2. Use BinMinSetCover to **generate "basic sparse solutions"** (with few rules).
3. **Apply dynamic programming** to optimize the support of these basic solutions.

Points 1. and 2. are collectively referred to in the paper as **ContMinSetCover**.

Point 3. is the **ContMaxSupport** algorithm.

ContMinSetCover and ContMaxSupport

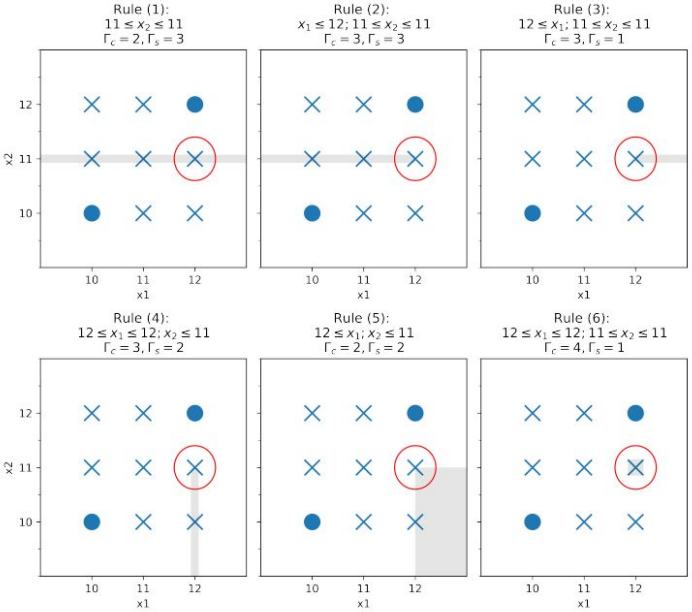


Figure 7: A simple example of "basic" rule-based summary-explanations for a continuous dataset.

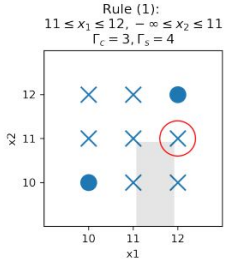


Figure 8: A simple example of a potentially optimal rule-based summary-explanation for a continuous dataset, found in the second step of our approach. We will later determine that this rule is not optimal, because a sparser rule with higher support exists.

Idea and visualization of ContMaxSupport

Algorithm 2 Algorithm `ContMaxSupport`

Input: γ (number of initial rules to extract), w_s and w_c (objective coefficients), $\mathbf{X}$ (data matrix), observation to explain $\mathbf{x}_e$, and predictions $\{h^{\text{global}}(\mathbf{x}_i)\}_{i=1}^{N}, h^{\text{global}}(\mathbf{x}_e)$.
Output: globally-consistent rule $\left(R, \tau, h^{\text{global}}(\mathbf{x}_e)\right)$.

1. Apply `ContMinSetCover` to extract γ rules with optimal sparsity (e.g., by running `ContMinSetCover` iteratively with additional cutting-planes that prohibit previous solutions).
2. Apply the DP formulation (12) to each of the extracted rules $h_{R,\tau,h^{\text{global}}(\mathbf{x}_e)}$ to increase their support.
3. Return the expanded rule whose objective value is maximal.

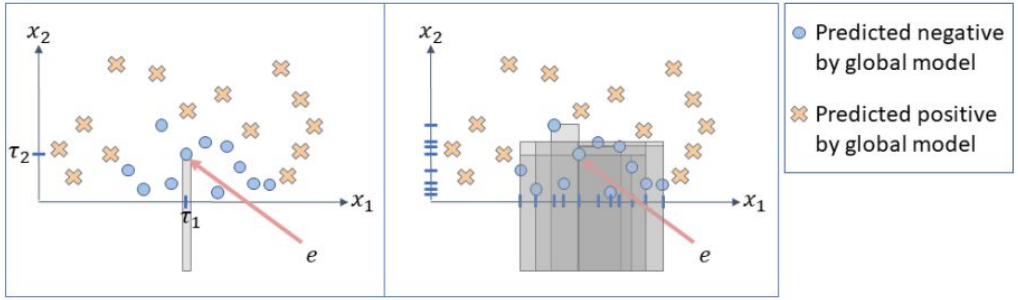


Figure 9: An illustration of basic rule and states of the corresponding DP. Left: the initial state of the DP, which is a summary-explanation for observation e (in gray). Right: states of the DP, each corresponding to a summary-explanation.

Numerical experiments

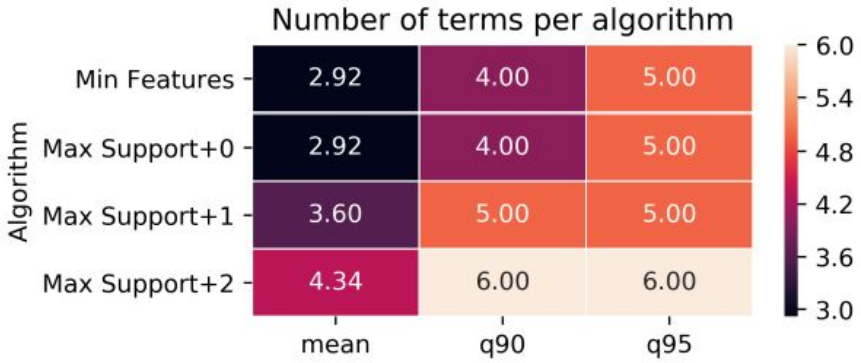
Evaluation on the FICO dataset which contains data about applications for home owner loans, **using 7 different models**. For the experiments it is preprocessed in 4 different ways:

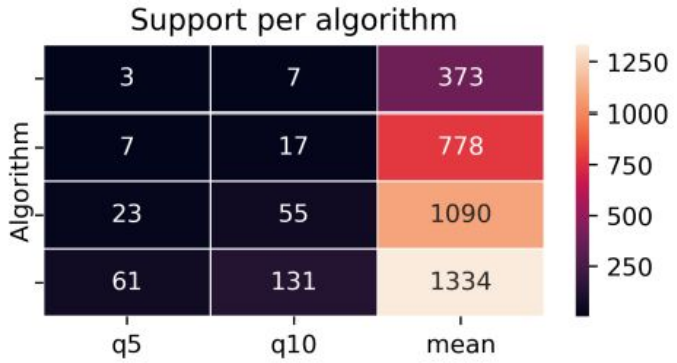
- **original** - no changes are made, missing values encoded as negative numbers
- **missing as binary** - for each feature add an indicator if this value is missing
- **quantiles** - replace the features with indicators of whether the observation is ≤ the quantile
- **manual** - manually split each variable into two groups

The first two methods yield **continuous** datasets, whereas the others yield **binary** datasets.

Results for binary datasets

1. The explanations on real datasets tend to be very **sparse** (require very few rules):



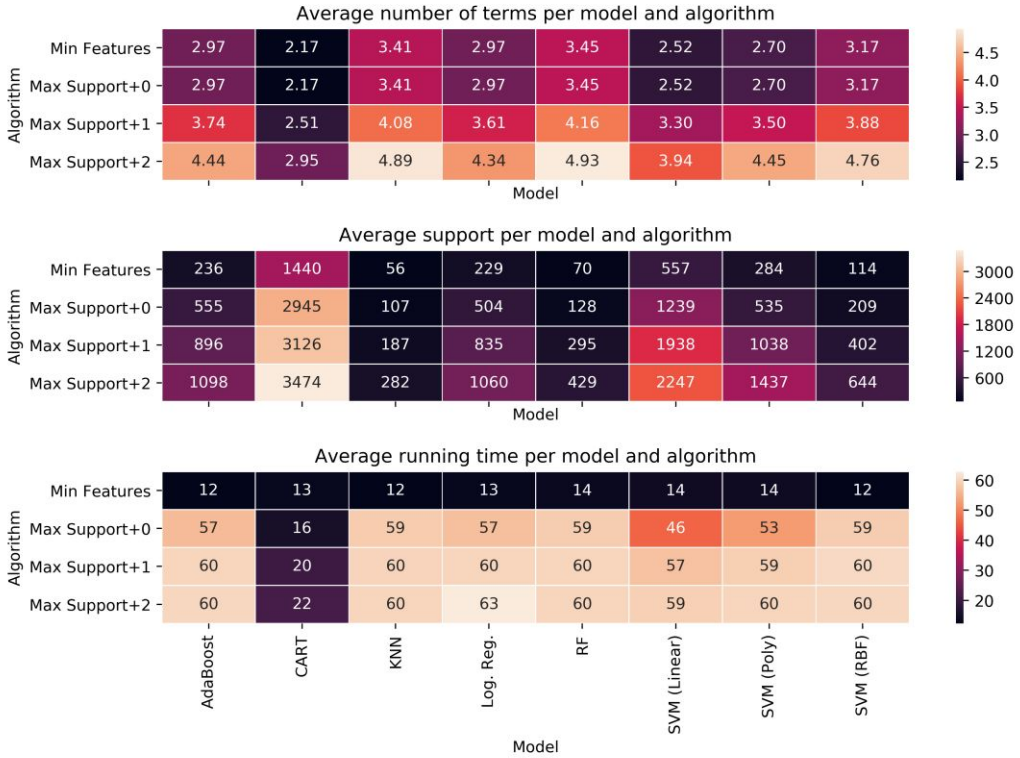


Results are averaged for all considered models and explanations for all observations

Results for binary datasets

2. The running times are **fast**.
3. **The rules are robust** to the underlying model.

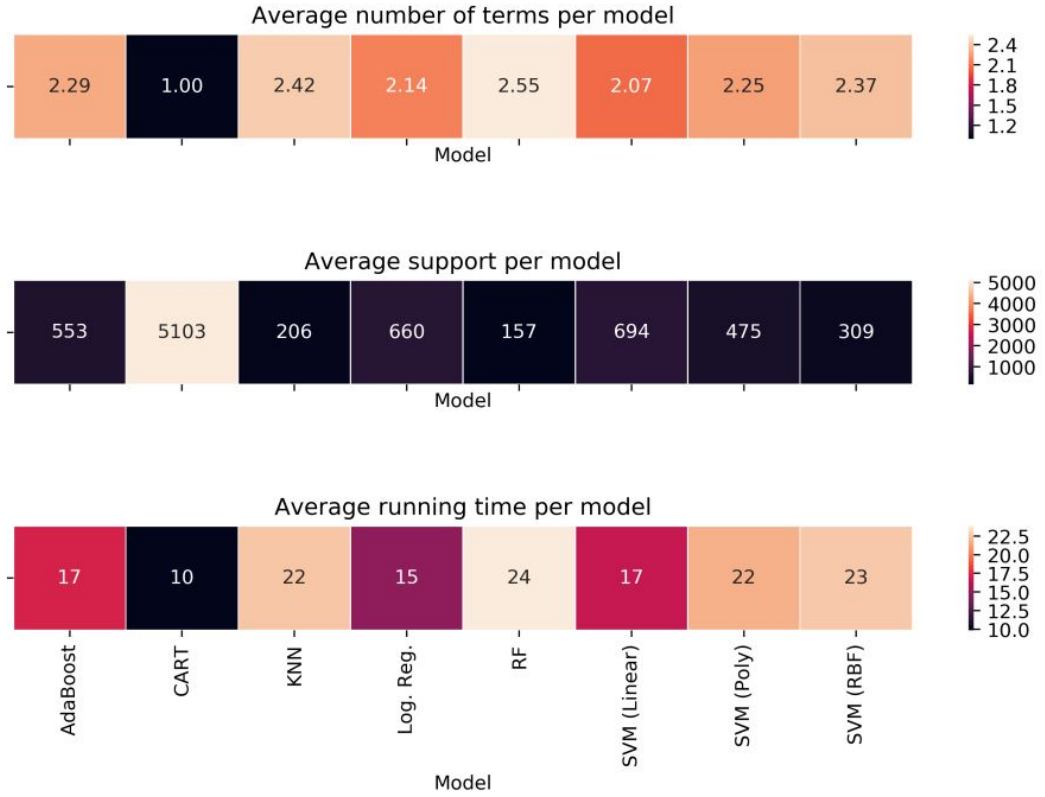
Each cell is an average of 9871 explanations (one for each observation)



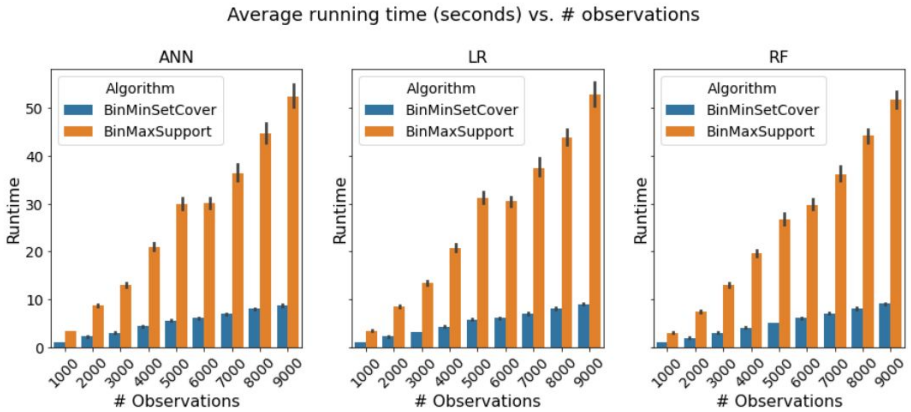
Results for continuous datasets

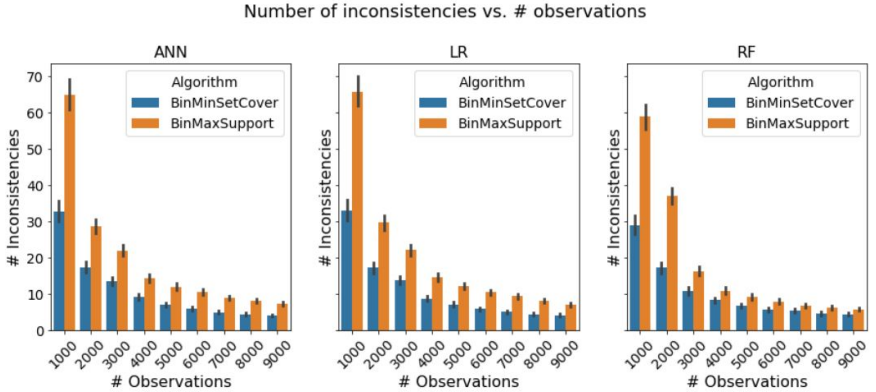
1. The explanations are **sparse**.
2. The **support is** quite **large**.
3. The explanations are **robust** to the underlying model.
4. The running time is quite **fast**.

Each cell is an average of 9871 explanations (one for each observation)



Additional robustness checks





Discussion points

- Runtime
- Privacy concerns (the rule-based explanations give insight into the data distribution)
- Misinterpretation of the result (the explanations do not imply causality)
- Necessity of interpretable features
- Stability of explanations
- Allowing for some inconsistency
- Generalization to new observations
- Summarizing global models
- Connections to counterfactuals

Thank you for your attention!