Computer Science > Machine Learning

# Lifting Interpretability-Performance Trade-off via Automated Feature Engineering

Alicja Gosiewska, Przemyslaw Biecek

*(Submitted on 11 Feb 2020)*

Complex black-box predictive models may have high performance, but lack of interpretability causes problems like lack of trust, lack of stability, sensitivity to concept drift. On the other hand, achieving satisfactory accuracy of interpretable models require more time-consuming work related to feature engineering. Can we train interpretable and accurate models, without timeless feature engineering? We propose a method that uses elastic black-boxes as surrogate models to create a simpler, less opaque, yet still accurate and interpretable glass-box models. New models are created on newly engineered features extracted with the help of a surrogate model. We supply the analysis by a large-scale benchmark on several tabular data sets from the OpenML database. There are two results 1) extracting information from complex models may improve the performance of linear models, 2) questioning a common myth that complex machine learning models outperform linear models.

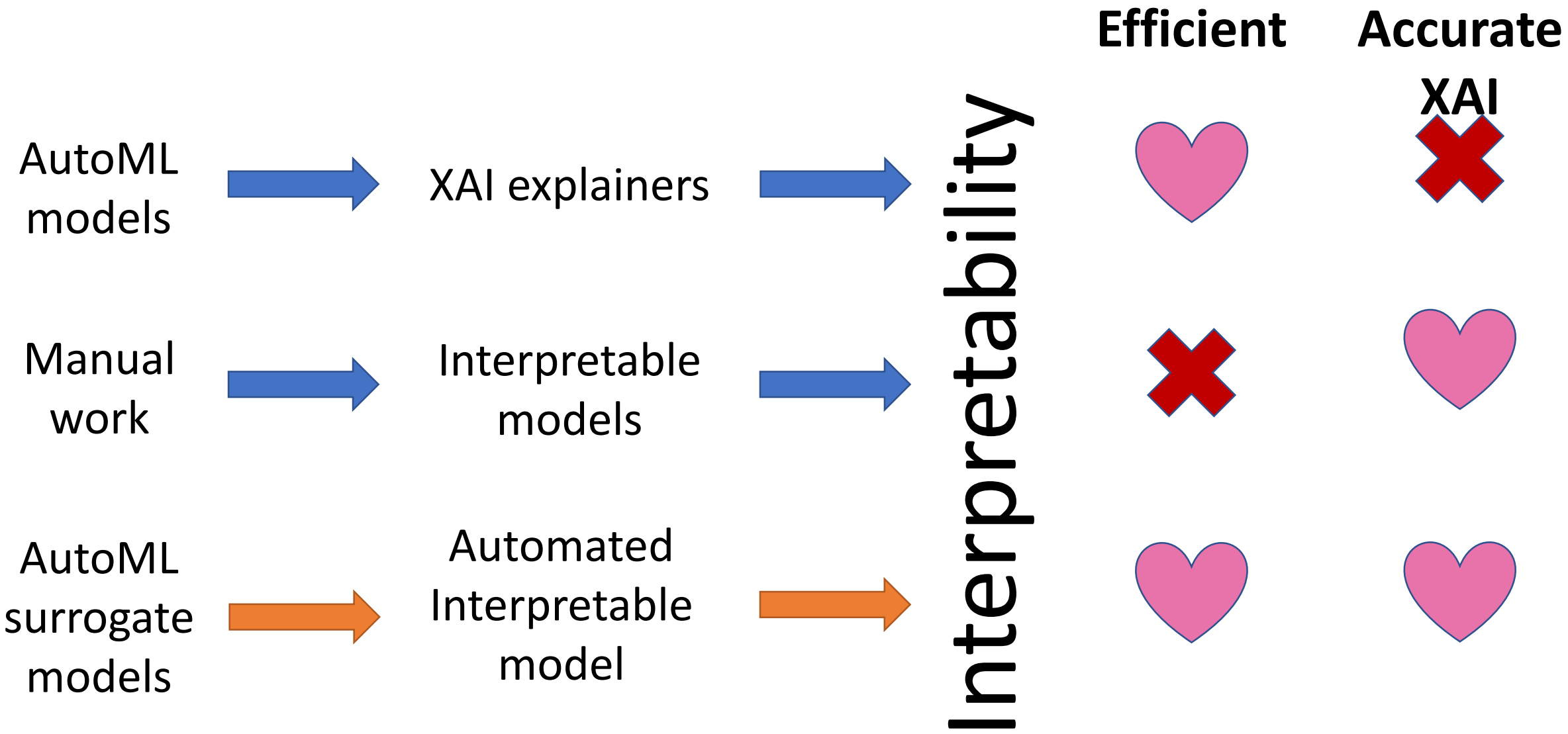


https://arxiv.org/abs/2002.04267

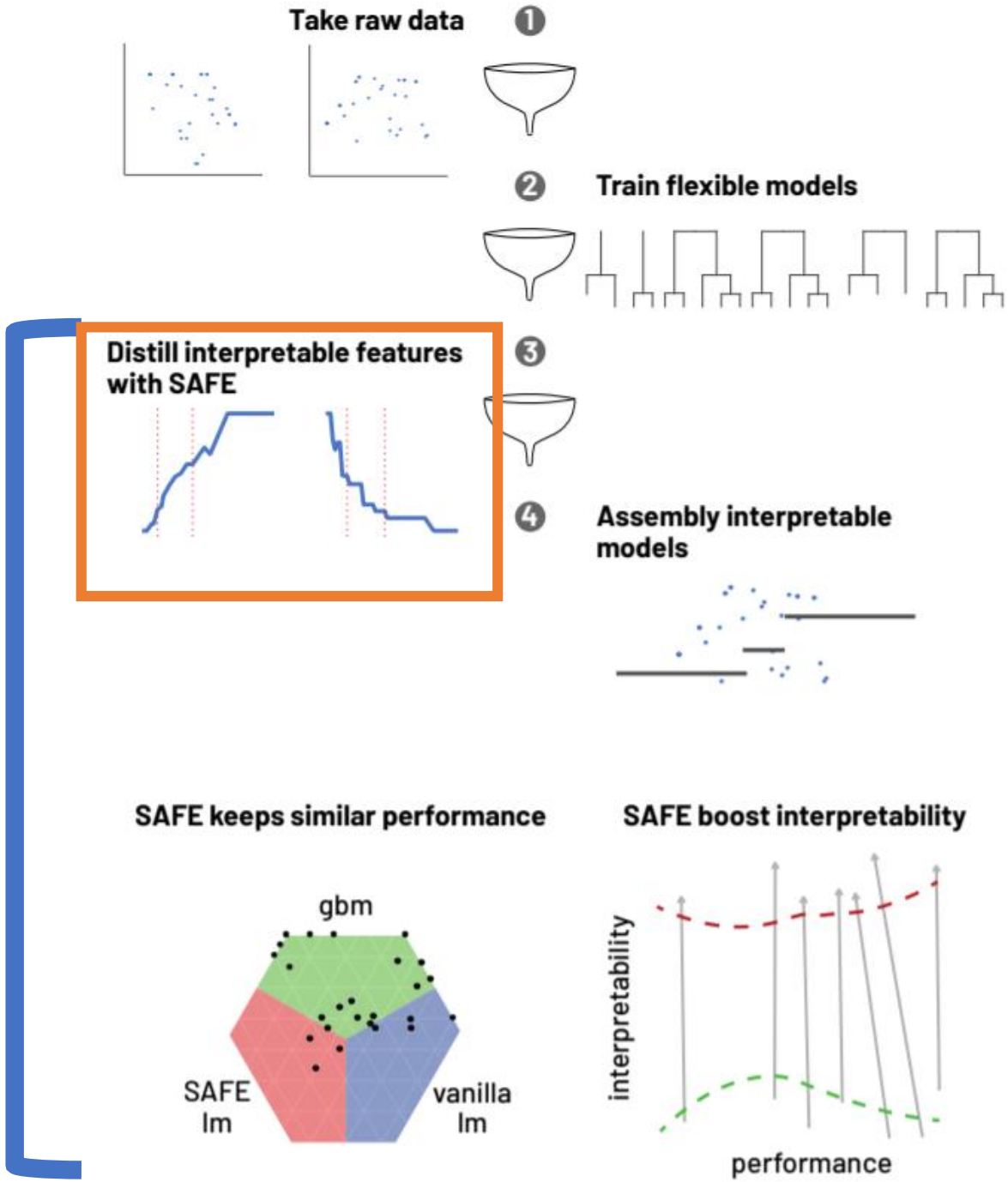
https://github.com/ModelOriented/rSAFE

https://github.com/ModelOriented/SAFE

# Paths of interpretability

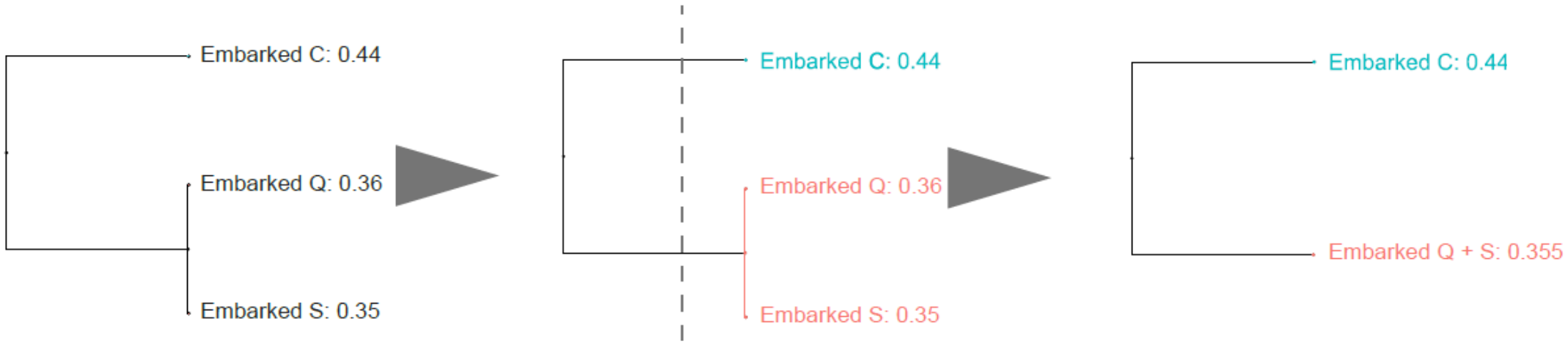


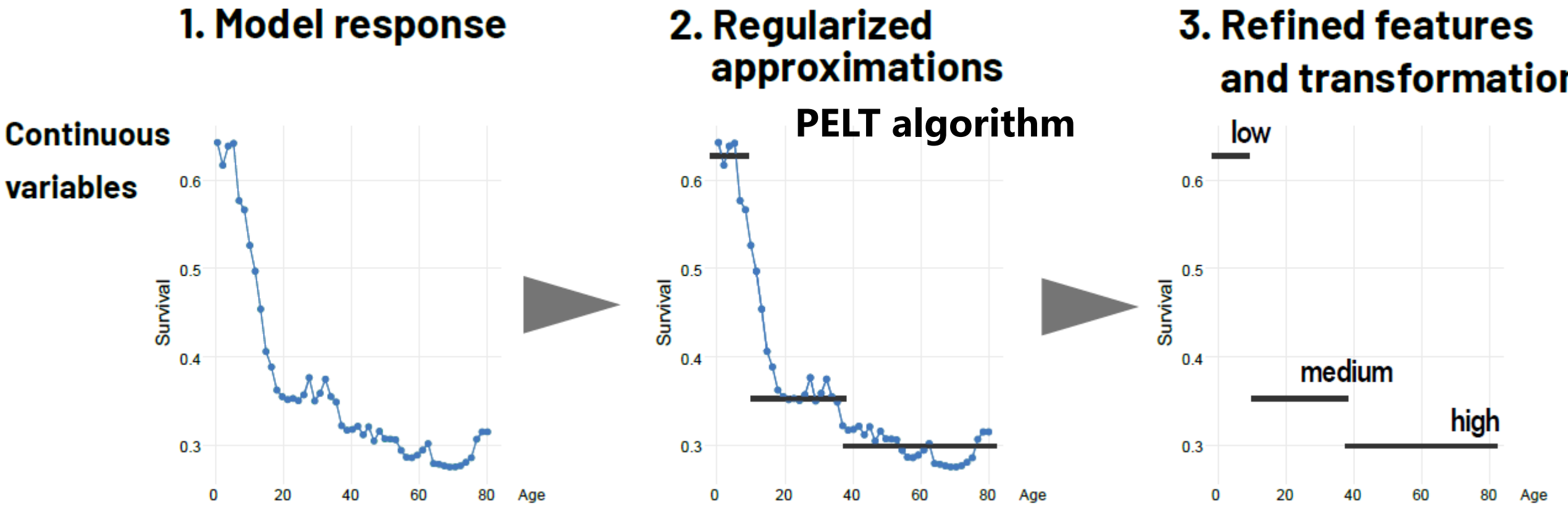
# Lifting Interpretability-Performance Trade-off via Automated Feature Engineering



## Categorical variables

## Hierarchical clustering






PELT Algorithm: Killick R, Fearnhead P, Eckley IA (2012) Optimal detection of changepoints with a linear computational cost, JASA 107(500), 1590–1598

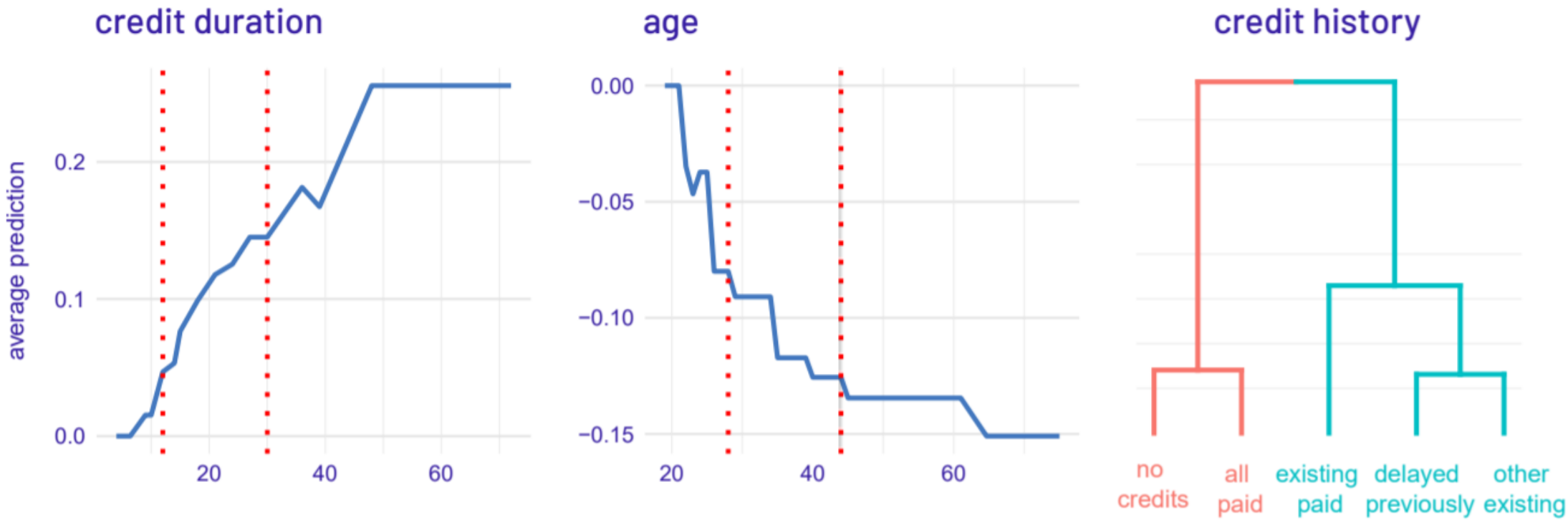



Figure 5: Example transformations of two continuous variables, credit duration and age, and a categorical variable credit history.

| | m2.price | construction.year | surface | floor | no.rooms | district |
|---|---|---|---|---|---|---|
| 1 | 5897 | 1953 | 25 | 3 | 1 | Srodmiescie |
| 2 | 1818 | 1992 | 143 | 9 | 5 | Bielany |
| 3 | 3643 | 1937 | 56 | 1 | 2 | Praga |
| 4 | 3517 | 1995 | 93 | 7 | 3 | Ochota |
| 5 | 3013 | 1992 | 144 | 6 | 5 | Mokotow |
| 6 | 5795 | 1926 | 61 | 6 | 2 | Srodmiescie |

**SAFE transformation**

| | m2.price | construction.year_new | surface_new | floor_new | no.rooms_new | district_new |
|---|---|---|---|---|---|---|
| 1 | 3542 | (1937, 1994] | (-Inf, 47] | (5, Inf) | (-Inf, 3] | Bemowo_Bielany_Praga_Ursus_Ursynow_Wola |
| 2 | 5631 | (1994, Inf) | (101, Inf) | (-Inf, 5] | (3, Inf) | Srodmiescie |
| 3 | 2989 | (1937, 1994] | (-Inf, 47] | (5, Inf) | (-Inf, 3] | Bemowo_Bielany_Praga_Ursus_Ursynow_Wola |
| 4 | 3822 | (1937, 1994] | (-Inf, 47] | (-Inf, 5] | (-Inf, 3] | Bemowo_Bielany_Praga_Ursus_Ursynow_Wola |
| 5 | 2337 | (1937, 1994] | (101, Inf) | (-Inf, 5] | (3, Inf) | Bemowo_Bielany_Praga_Ursus_Ursynow_Wola |
| 6 | 3381 | (1937, 1994] | (47, 101] | (5, Inf) | (-Inf, 3] | Mokotow_Ochota_Zoliborz |

# Benchmark

Table 2: Mean AUC of models followed by standard deviation, calculated from 10 train/test splits defined for each data set in the OpenML database. The highest values of AUC for each data set are bolded.

| dataset (OML task) | vanilla logistic regression | gbm default | SAFE gbm default | gbm tuned | SAFE gbm tuned | svm | SAFE svm |
|---|---|---|---|---|---|---|---|
| credit-g (31) | **0.79+-0.04** | 0.78+-0.04 | 0.77+-0.04 | 0.78+-0.05 | 0.77+-0.03 | **0.79+-0.04** | 0.73+-0.05 |
| diabetes (37) | 0.83+-0.06 | 0.83+-0.04 | **0.84+-0.04** | **0.84+-0.04** | 0.83+-0.04 | 0.83+-0.05 | 0.83+-0.04 |
| spambase (43) | 0.97+-0.01 | **0.98+-0.01** | **0.98+-0.01** | **0.98+-0.01** | **0.98+-0.01** | **0.98+-0.01** | **0.98+-0.01** |
| tic-tac-toe (49) | **1.00+-0** | 0.81+-0.03 | 0.82+-0.04 | **1.00+-0** | 0.74+-0.05 | **1.00+-0** | 0.75+-0.05 |
| electricity (219) | 0.75+-0.08 | 0.86+-0.01 | 0.86+-0.01 | **0.92+-0** | 0.86+-0.01 | 0.88+-0 | 0.84+-0.01 |
| scene (3485) | 0.96+-0.02 | **0.98+-0.02** | 0.87+-0.03 | **0.98+-0.02** | 0.77+-0.02 | 0.94+-0.02 | 0.71+-0.03 |
| monks-problems-1 (3492) | 0.70+-0.07 | 0.69+-0.06 | 0.70+-0.06 | 0.72+-0.06 | 0.72+-0.08 | **1+-0** | 0.71+-0.08 |
| monks-problems-2 (3493) | 0.54+-0.10 | 0.54+-0.10 | 0.55+-0.11 | 0.53+-0.09 | 0.52+-0.07 | **0.65+-0.06** | 0.56+-0.10 |
| monks-problems-3 (3494) | **0.99+-0.02** | 0.98+-0.03 | **0.99+-0.02** | **0.99+-0.02** | **0.99+-0.02** | 0.98+-0.03 | **0.99+-0.02** |
| gina_agnostic (3891) | 0.79+-0.02 | 0.92+-0.02 | 0.78+-0.03 | 0.94+-0.02 | 0.80+-0.03 | **0.96+-0.01** | 0.80+-0.03 |
| mozilla4 (3899) | 0.89+-0.01 | 0.96+-0.01 | 0.90+-0.02 | **0.97+-0.01** | 0.89+-0.02 | 0.93+-0.01 | 0.91+-0.01 |
| pc4 (3902) | 0.92+-0.03 | 0.93+-0.02 | 0.89+-0.03 | **0.94+-0.02** | 0.89+-0.03 | 0.90+-0.02 | 0.84+-0.05 |
| pc3 (3903) | **0.82+-0.06** | **0.82+-0.03** | 0.78+-0.06 | **0.82+-0.04** | 0.79+-0.07 | 0.72+-0.08 | 0.79+-0.06 |
| kc2 (3913) | 0.82+-0.12 | **0.85+-0.09** | 0.82+-0.09 | 0.84+-0.11 | 0.83+-0.11 | 0.78+-0.1 | 0.81+-0.12 |
| kc1 (3917) | **0.80+-0.03** | **0.80+-0.04** | 0.79+-0.04 | **0.80+-0.04** | 0.79+-0.04 | 0.74+-0.06 | 0.79+-0.03 |
| pc1 (3918) | 0.81+-0.07 | 0.82+-0.06 | 0.80+-0.07 | **0.83+-0.06** | 0.81+-0.09 | 0.78+-0.05 | 0.80+-0.08 |
| MagicTelescope (3954) | **1.00+-0** | **1.00+-0** | 0.99+-0 | **1.00+-0** | **1.00+-0** | **1.00+-0** | **1.00+-0** |
| wdbc (9946) | 0.95+-0.03 | **0.99+-0.01** | 0.97+-0.03 | **0.99+-0.01** | **0.99+-0.01** | **0.99+-0.01** | 0.96+-0.03 |
| phoneme (9952) | 0.81+-0.02 | 0.87+-0.01 | 0.87+-0.02 | 0.90+-0.01 | 0.88+-0.01 | **0.91+-0.01** | 0.86+-0.01 |
| qsar-biodeg (9957) | 0.92+-0.03 | 0.91+-0.03 | 0.91+-0.03 | 0.92+-0.03 | 0.91+-0.03 | **0.93+-0.03** | 0.90+-0.04 |
| hill-valley (9970) | 0.59+-0.04 | 0.53+-0.04 | 0.55+-0.04 | **0.60+-0.06** | 0.58+-0.06 | 0.54+-0.07 | 0.53+-0.03 |
| ilpd (9971) | **0.75+-0.07** | 0.73+-0.06 | 0.73+-0.05 | 0.73+-0.05 | 0.73+-0.07 | 0.66+-0.08 | 0.73+-0.08 |
| madelon (9976) | 0.59+-0.04 | **0.69+-0.03** | 0.63+-0.03 | 0.68+-0.03 | 0.63+-0.04 | 0.62+-0.04 | 0.53+-0.01 |
| ozone-level-8hr (9978) | **0.90+-0.04** | 0.89+-0.04 | 0.89+-0.04 | **0.90+-0.03** | 0.88+-0.04 | **0.90+-0.04** | 0.83+-0.04 |
| climate-model-simulation-crashes (9980) | **0.85+-0.1** | 0.82+-0.15 | 0.77+-0.1 | 0.81+-0.14 | 0.81+-0.11 | **0.85+-0.07** | 0.77+-0.08 |
| eeg-eye-state (9983) | 0.68+-0.01 | 0.78+-0.01 | 0.77+-0.01 | 0.85+-0.01 | 0.79+-0.01 | **0.88+-0.03** | 0.77+-0.01 |
| banknote-authentication (10093) | **1.00+-0** | 0.99+-0.01 | 0.99+-0.01 | **1.00+-0** | **1.00+-0** | **1.00+-0** | 0.99+-0.01 |
| blood-transfusion-service-center (10101) | **0.75+-0.05** | **0.75+-0.05** | 0.74+-0.05 | **0.75+-0.05** | 0.74+-0.05 | 0.69+-0.05 | 0.71+-0.04 |
| bank-marketing (14965) | 0.91+-0.01 | 0.90+-0.01 | 0.89+-0.01 | **0.92+-0.01** | 0.88+-0.01 | 0.90+-0.01 | 0.89+-0.01 |
| PhishingWebsites (34537) | **0.99+-0** | 0.98+-0 | 0.98+-0 | **0.99+-0** | 0.98+-0 | **0.99+-0** | 0.98+-0 |

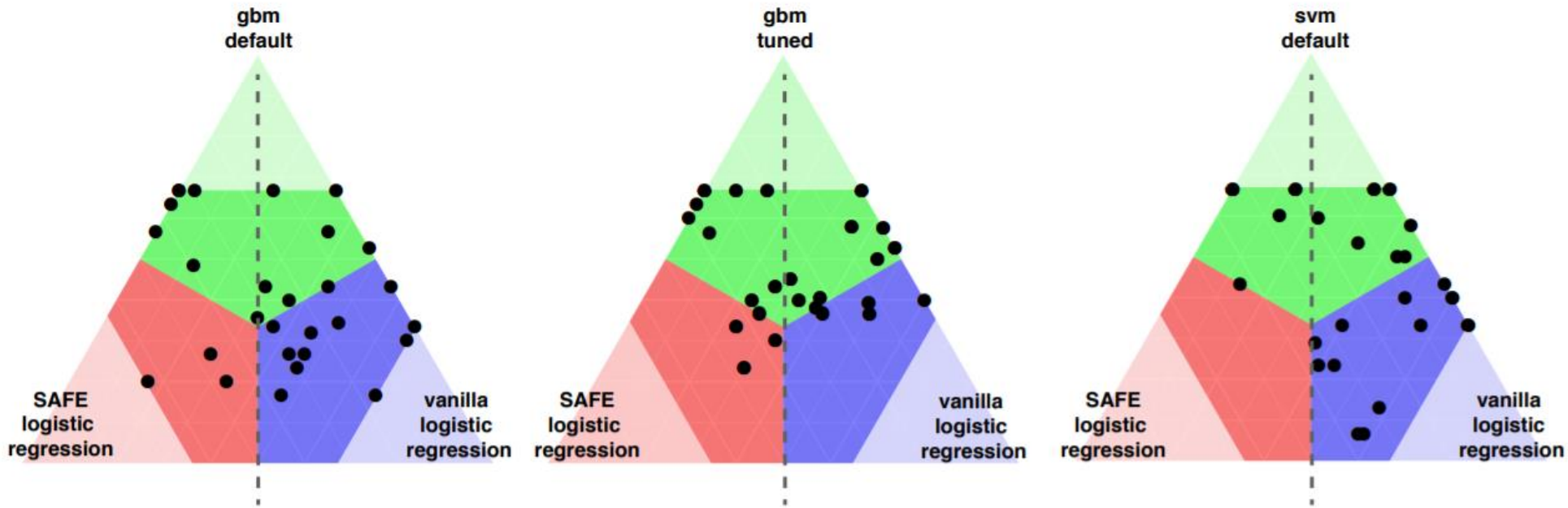



Figure 3: Ternary plots of AUC measures. One dot corresponds to models' performances on one data set. The position in the triangle is composed of AUC values of vanilla logistic regression, surrogate model, and refined logistic regression. Dots in the green area are data sets for which, on average, surrogate model was the best. Dots in the red area indicates data sets for which models refined with the SAFE method were the best, the blue area contains dots for which vanilla logistic regression was the best. On the left side of the vertical dashed line are data sets for which SAFE-based logistic regression models where on average better than vanilla logistic regression. The area marked by more saturated color shows a range of dots' possible appearance area. Dots cannot reach corner areas because their positions are calculated based on positions in AUC ranking among three models (baseline, surrogate, and refined). It means that for a split in a data set, the best model gains 2 points, second gains 1 point, third gains 0 points. After averaging over 10 splits we obtain trinomial vector with averaged scores for three models. If the model would win on all splits, it would gain $\frac{2}{3}$ of the total sum of points, thereby not all parts of the triangle are reachable. Models that gain $\frac{2}{3}$ of the total sum of points lie on the one-colored edges of the hexagon.

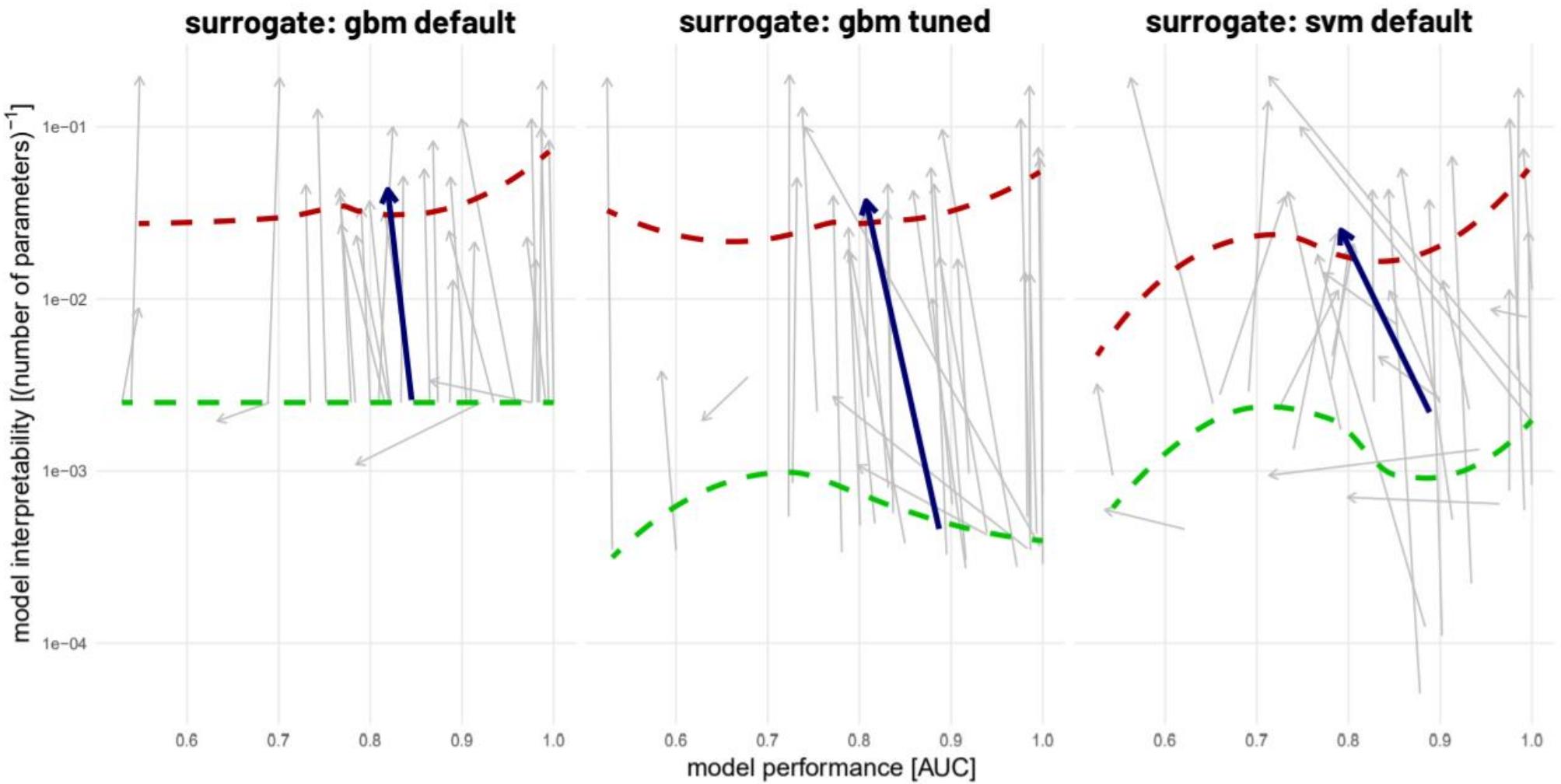



Figure 4: The interpretability-performance trade-off. The beginnings of grey arrows mark complex surrogate models' performances and their interpretability levels, the arrowheads mark SAFE-based refined models' performances and their interpretability levels. Therefore, grey arrows illustrate interpretability-performance shifts for data sets when using SAFE. The dark blue arrows shows medians. The Green dashed lines are interpretability trends for surrogate models, the red dashed lines are interpretability trends for SAFE-based refined models. Vertical offsets between these lines shows that SAFE lifted the interpretability-performance trade-off.