Selecting Data Assets in Data Marketplaces – Leveraging Machine Learning and Explainable AI for Value Quantification

Dominik Martin, Daniel Heinz, Moritz Glauner, Niklas Kühl

Business & Information Systems Engineering (2025)

Appendix (available online via http://link.springer.com)

A Appendix

A.1 Static Data-based Quantification Methods

Feature selection methods used in our proposed artifact include the following filter and wrapper methods:

Variance-based feature selection ranks features according to their variance. The idea behind this filter method is to eliminate features that mainly consist of noise (Bommert et al. 2020). To avoid biased scores, if features are scaled differently, we normalize the variance using the mean of the respective feature.

The *ANOVA-based feature selection* filter method makes use of the analysis of variance (ANOVA), a statistical test that evaluates if the means of different groups are significantly different from each other. In the context of feature selection, ANOVA is applied to each feature against the target variable. The dataset is categorized into groups based on the values of the feature, and the mean value of the target variable is calculated for each group. The ANOVA test then checks if these means are significantly different. The result of the ANOVA test, typically the F-statistic, is utilized to rank the features, with a higher F-statistic indicating a more predictive feature (Bommert et al. 2020).

Mutual information feature selection is another filtering method that scores features according to their degree of dependence or correlation with the target variable. The mutual information score is an entropy-based measure that quantifies the amount of information obtained about one variable through observing another variable.

In the context of feature selection, for each feature, a mutual information score is computed with respect to the target variable. The higher the mutual information score, the more the feature and target variable depend on each other, and the more likely the feature will be useful in predicting the target variable.

This method can handle both numerical and categorical data and does not assume a linear relationship between variables, making it versatile. However, it requires a good estimation of the probability distribution of the variables, which can be challenging for high-dimensional data or small datasets (Bommert et al. 2020; Brown et al. 2012; Hall 2000).

Sequential feature selection (SFS) is a greedy search-based heuristic that starts with an empty set of features and sequentially adds the most valuable feature at each step, which is determined by the marginal improvement it provides to a model's performance. Thus, this wrapper method can use any model but is computationally expensive (Pudil et al. 1994).

In each iteration, the algorithm evaluates the performance of a model with an additional feature from the remaining pool of features that are not yet selected. The feature which results in the highest performance increase is added to the set.

SFS, as it is a greedy algorithm, it may not find the optimal feature subset if there is a need to remove features added in earlier steps (Ferri et al. 1994; Aha and Bankert 1995).

We do not use embedded feature selection methods because they rely on model-inherent functionalities and thus cannot be used independently as desired in DR4.

A.2 Static Model-based Quantification Methods

Gloabl XAI methods used in our proposed artifact include:

Permutation importance quantifies the importance of a feature by measuring the increase of the prediction error of a model after permuting the features' values (Altmann et al. 2010; Molnar 2019). An implementation of the method is provided through the inspection module within the scikit-learn library in Python.

Partial dependence plots (PDP) analyze the marginal effect a feature has on the prediction of a ML model (Friedman 2001). Theoretically, the marginal effect can also be calculated for more than one feature, but this was not applied here. An implementation of the method is provided through the scikit-learn library in Python as well.

Accumulated local effects (ALE) describes how a given feature affects the prediction of a ML model

on average. It can be seen as a derivative to the Partial Dependence Plots mentioned above. An implementation of the method is provided through the ALIBI Explain Library (Klaise et al. 2021).

ELI5 model inspection is not a model-agnostic method but a library collecting numerous model-specific inspection approaches. Therefore, it strictly does not fulfill the DRs. However, as the library enables the inspection of a wide range of models, it achieves a similar effect as model-agnostic approaches, despite being model specific.

Local XAI methods used in our proposed artifact include:

*Shapley values*are a game theoretical construct used to calculate the contribution and thus the payout of every participant to a coalition (Shapley 1953). This construct can be used in order to quantify the contribution of a feature to a prediction in a ML model (Molnar 2019). Many different approaches exist to estimate the Shapley values. In this paper, we use SHAP, which is available through a proprietary package (Lundberg and Lee 2017).

LIME is an acronym for "Local interpretable model-agnostic explanations". The basic idea is to use interpretable surrogate models to approximate a black-box model. These surrogate models are trained locally for every instance using a dataset created from perturbations of the sample to be explained and the predicted target for these perturbations. The result claims to be valid only for the sample inspected. This is called local fidelity (Ribeiro et al. 2016b; Molnar 2019).

A.3 Dataset and Algorithm Details

This section provides detailed descriptions of the datasets, algorithms, and parameter settings used in our study, which are essential for replicating and evaluating the research presented. Tables included herein enumerate the datasets, performance estimation and modeling aspects and, algorithm parameters employed, aligning with the goals of the 'Supervised Machine Learning Reportcard (SMLR)' framework (Kühl et al. 2021) to enhance documentation standards in IS research.

Table 7: Adult Dataset.

Aspect	**Details**
Problem Statement	Predict if income exceeds $50K/year based on census data.
Data Gathering	Collected from the 1994 U.S. Census database.
Data Distribution	48,842 instances, 14 features expanded to 106 through one-hot encoding.
Data Quality	Well-documented and widely used for benchmarking classification models.

Table 8: Auto MPG.

Aspect	**Details**
Problem Statement	Predict the fuel consumption of vehicles based on various characteristics.
Data Gathering	Data sourced from 1970s car specifications.
Data Distribution	Contains 398 instances and 7 features.
Data Quality	Includes some missing values; generally clean and used for regression tasks.

Table 9: Boston Housing.

Aspect	**Details**
Problem Statement	Predict the median value of homes in various Boston districts.
Data Gathering	Derived from information collected by the U.S. Census Service.
Data Distribution	Comprises 506 instances, each with 14 features.
Data Quality	High-quality, though some ethical concerns have been raised about a feature related to racial proportions.

Table 10: Breast Cancer Wisconsin.

Aspect	Details
Problem Statement	Predict whether a breast mass is malignant based on image-derived features.
Data Gathering	Features computed from digitized images of fine needle aspirates of breast masses.
Data Distribution	Includes 569 instances with 30 features.
Data Quality	High-quality with well-documented feature set.

Table 11: California Housing.

Aspect	Details
Problem Statement	Predict house prices based on socioeconomic data.
Data Gathering	Data from the 1990 California census.
Data Distribution	Features 20,640 instances with 8 features.
Data Quality	Clean dataset, frequently used for regression.

Table 12: Diabetes.

Aspect	Details
Problem Statement	Predict the progression of diabetes in patients one year after baseline.
Data Gathering	Collected from a diabetes study.
Data Distribution	Contains 442 instances and 10 features.
Data Quality	Generally clean and used in regression tasks.

Table 13: Gas Turbine (CO2 and NOx emissions).

Aspect	Details
Problem Statement	Predict emissions from gas turbines.
Data Gathering	Data collected from gas turbines operating in Turkey.
Data Distribution	Each dataset (CO2 and NOx) has 36,733 instances and 11 features.
Data Quality	Includes continuous operational parameters, well-suited for regression.

Table 14: Heart Disease.

Aspect	Details
Problem Statement	Predict the presence of heart disease in patients.
Data Gathering	Compiled from several different healthcare studies.
Data Distribution	Consists of 303 instances, 14 original features expanded to 26 with encoding.
Data Quality	Widely used in binary classification tasks, high relevance and reliability.

Table 15: Wine.

Aspect	Details
Problem Statement	Classify the origin of wine based on physicochemical tests.
Data Gathering	Data derived from a chemical analysis of wines grown in a specific area of Italy.
Data Distribution	Includes 178 instances with 13 features.
Data Quality	Highly specific dataset, used for multiclass classification tasks.

A.4 Evaluation

This subsection presents additional insights into the technical evaluation described in Section 5. We provide cost-based evaluation plots covering all four scenarios for all the datasets evaluated (cf. Table 3). As discussed in Section 5.3, the experiments show slight performance differences between the various

Table 16: Industry.

Aspect	Details
Problem Statement	Classification of hydraulic seal conditions.
Data Gathering	Data gathered from various sensors on a test bench including temperatures, pressures, motion, friction, vibration, acoustics, room temperature, and humidity.
Data Distribution	36 distinct sensor readings preprocessed into 545 features.
Data Quality	High-quality data without missing values.

Table 17: Performance estimation and modeling aspects.

Parameter optimization	None
Data split	Training data set 80%, Evaluation data set 20%.
Sampling/Data augmentation	None
Performance metric	R2-score for regression tasks F1 score for classification tasks Area under the curve (AUC)

Table 18: Algorithm parameters - Random Forest

Number of estimators	100
Max depth	None
Min samples split	2
Min samples leaf	1
Performance evaluation	cf. Table 4

Table 19: Algorithm parameters - XGBoost

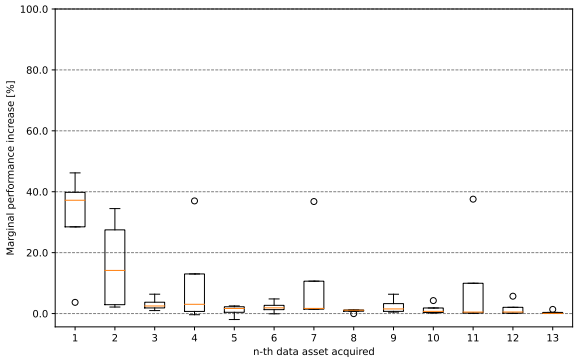
Learning rate	0.3
Gamma	0
Max depth	6
Min child weight	1
Performance evaluation	cf. Table 4

methods applied on different datasets. However, overall, in particular, some methods like SFS, mutual information (static data-based), SHAP and LIME (dynamic data-and model-based), and SHAP (global), PDP, and permutation importance perform consistently well.

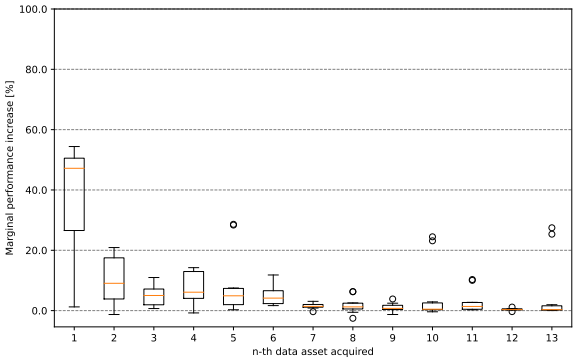
A.5 Nascent Design Knowledge for DSQM

This section provides detailed explanations of the nascent design knowledge for the DQSM, as outlined in Table 6 based on Jones and Gregor (2007). The key components and their descriptions are elaborated below:

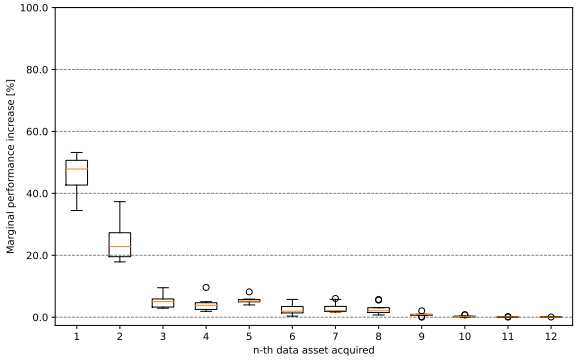
- **Purpose and Scope:** The DQSM design theory aims to provide prescriptive knowledge for creating mechanisms that enable data marketplaces to quantify and select data assets. By addressing challenges such as information asymmetry, variability in data value, and the need for context-sensitive decision-making, the DQSM empowers data consumers while preserving the interests of data providers.
- **Key Constructs:** Two core constructs (Offermann et al. 2010) underpin the DQSM:
  - *Level of Dependency:* This construct differentiates between *model-based* approaches, which assess data contributions relative to an existing model, and *data-based* approaches, which focus on the intrinsic properties of data assets.



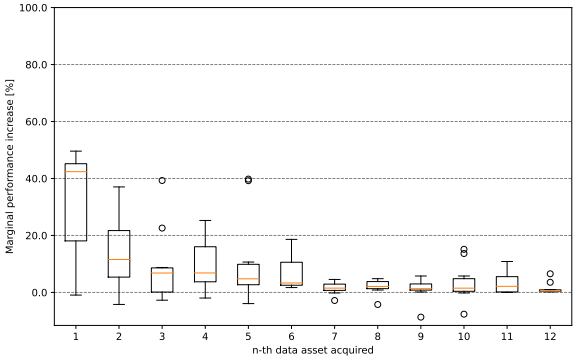
(a) Static data-based



(b) Dynamic data-based

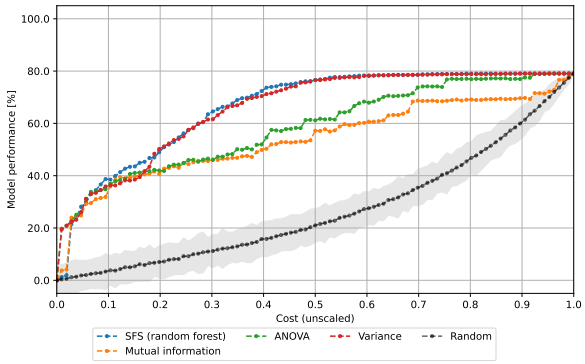


(c) Static model-based

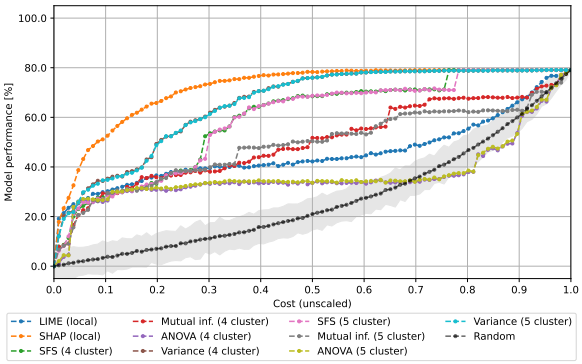


(d) Dynamic model-based

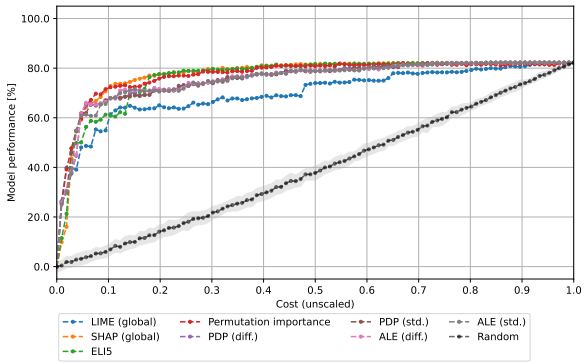
Figure 15: Marginal performance increase per n-th additionally selected data asset on Boston Housing (Harrison and Rubinfeld 1978) dataset. The graphs illustrate the distributions covering all presented methods except for the badly performing variance (static data-based) as well as the variance and ANOVA-based clustering approaches (dynamic data- and model-based).



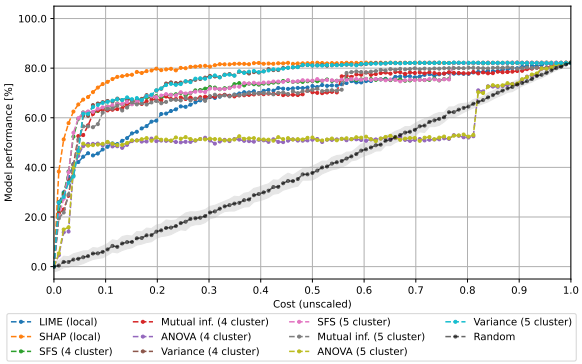
(a) Static data-based



(b) Dynamic data-based

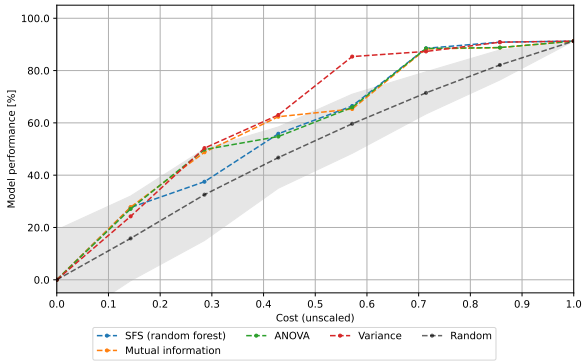


(c) Static model-based

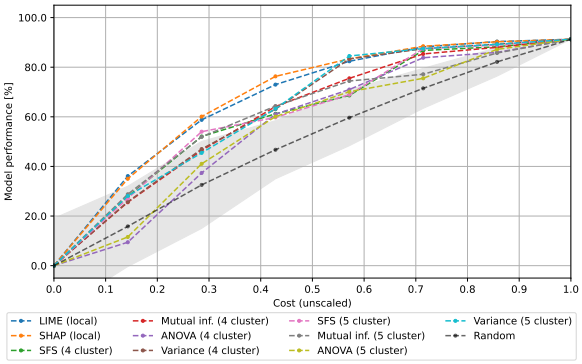


(d) Dynamic model-based

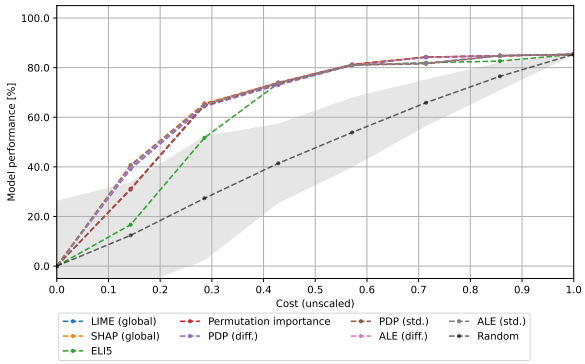
Figure 16: Results of cost-based evaluation on Adult (Becker and Kohavi 1996) dataset.



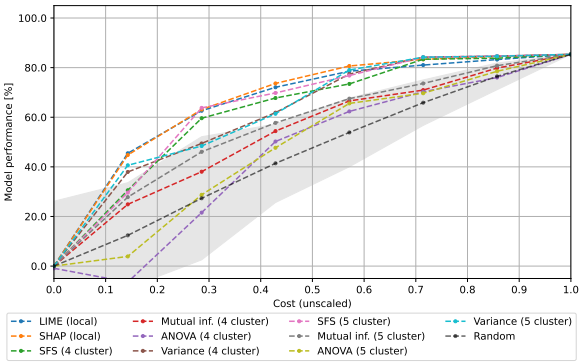
(a) Static data-based



(b) Dynamic data-based

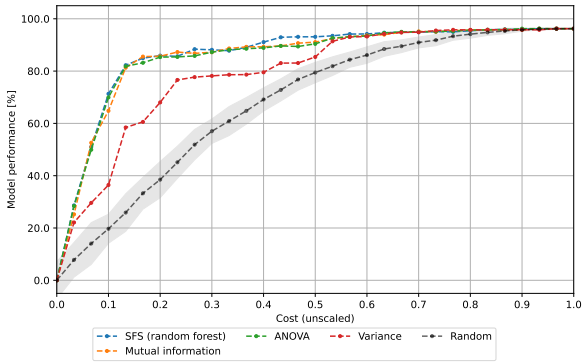


(c) Static model-based

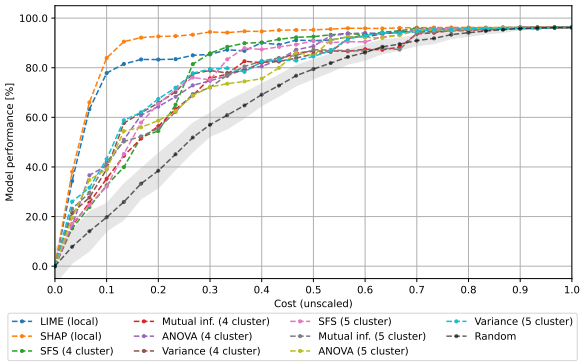


(d) Dynamic model-based

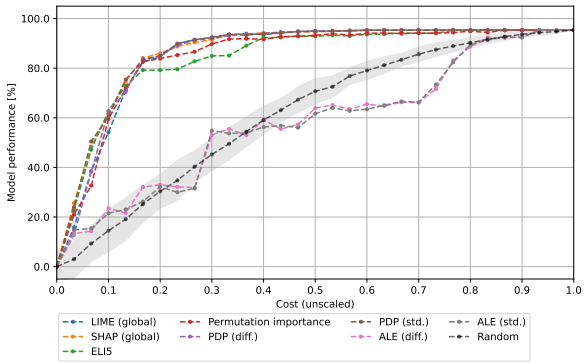
Figure 17: Results of cost-based evaluation on Auto MPG (Quinlan 1993) dataset.



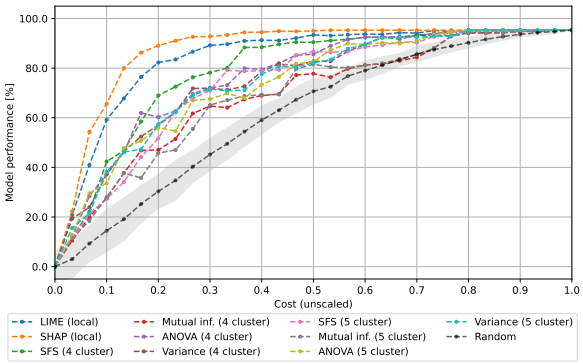
(a) Static data-based



(b) Dynamic data-based

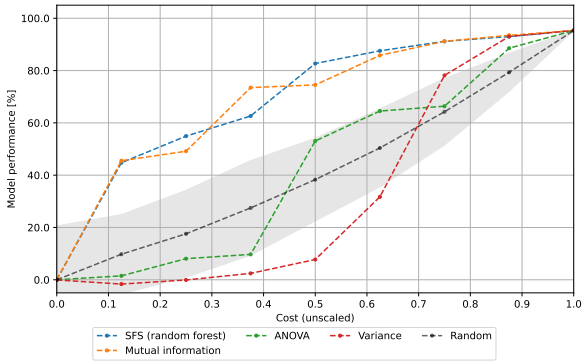


(c) Static model-based

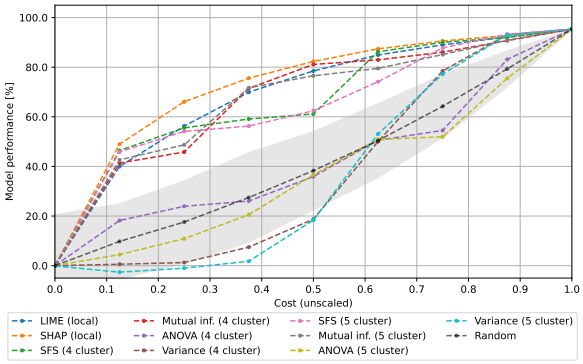


(d) Dynamic model-based

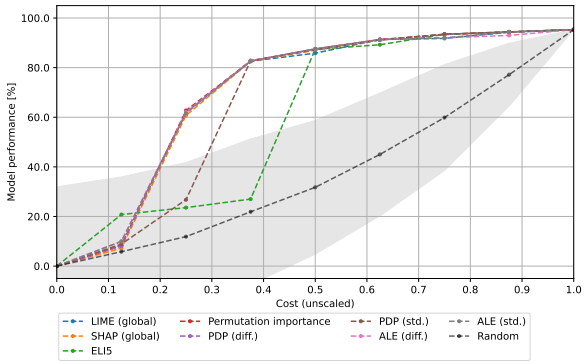
Figure 18: Results of cost-based evaluation on Breast Cancer (Wolberg et al. 1995) dataset.



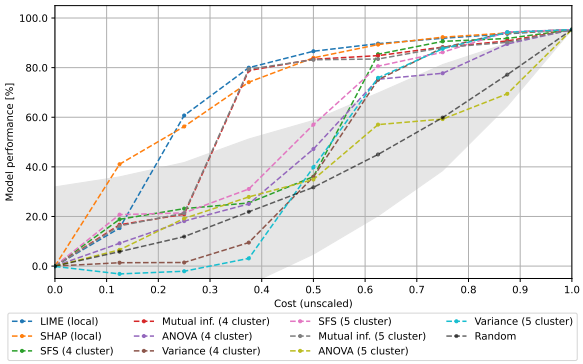
(a) Static data-based



(b) Dynamic data-based

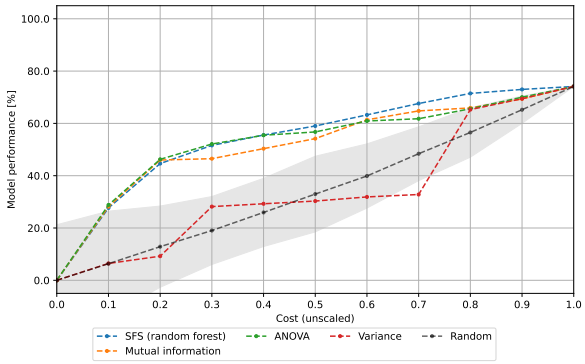


(c) Static model-based

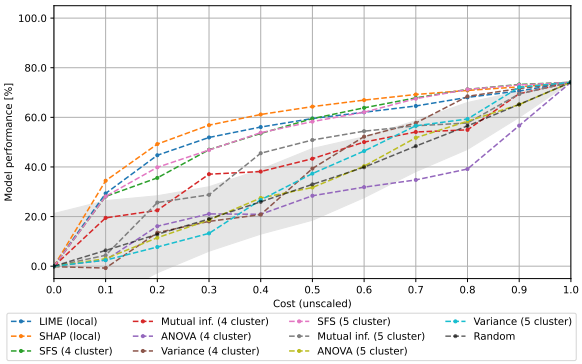


(d) Dynamic model-based

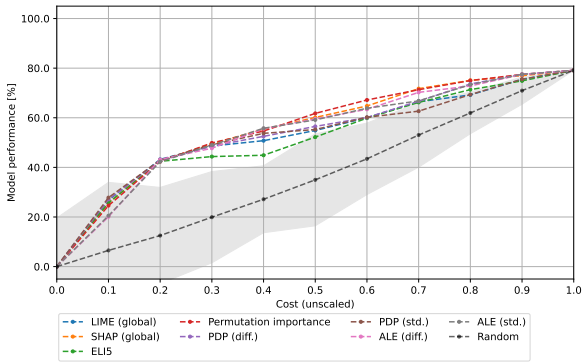
Figure 19: Results of cost-based evaluation on California Housing (Pace and Barry 1997) dataset.



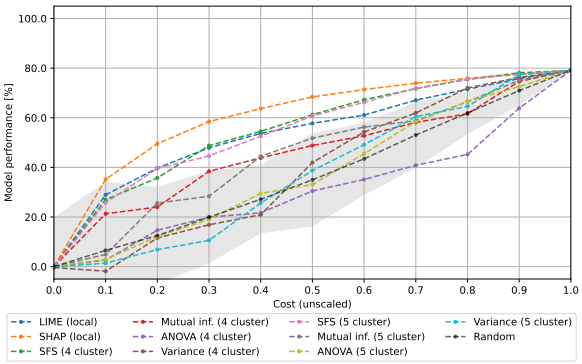
(a) Static data-based



(b) Dynamic data-based

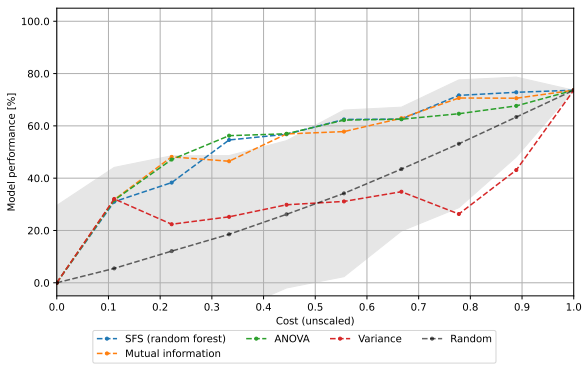


(c) Static model-based

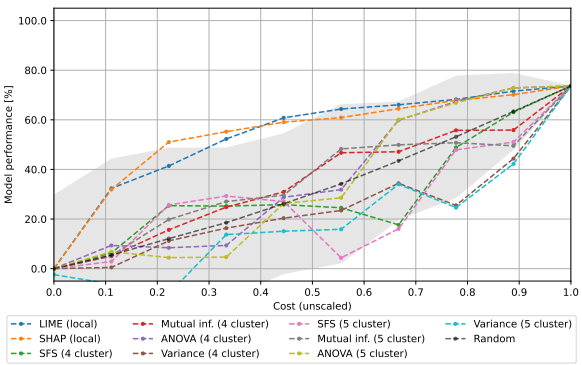


(d) Dynamic model-based

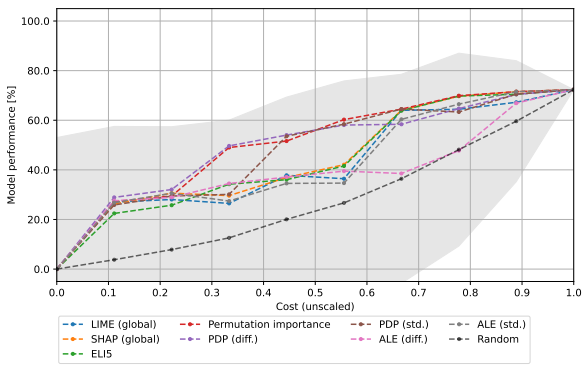
Figure 20: Results of cost-based evaluation on Diabetes (Kahn 1994) dataset.



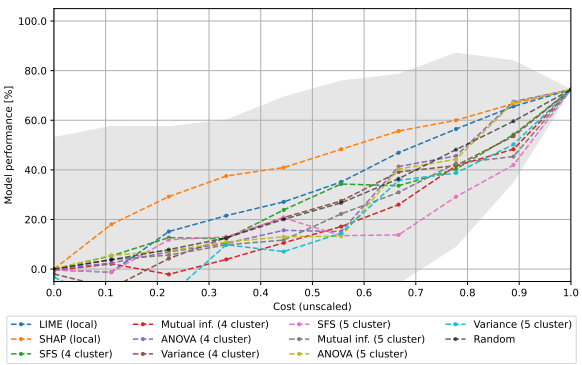
(a) Static data-based



(b) Dynamic data-based

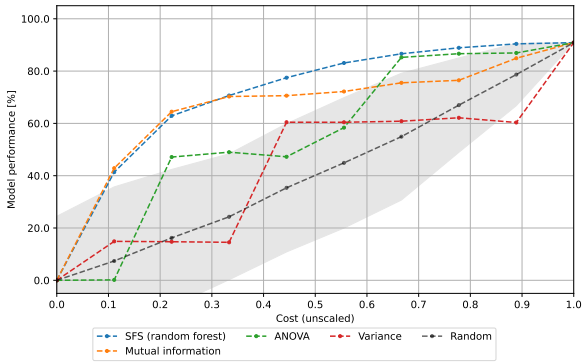


(c) Static model-based

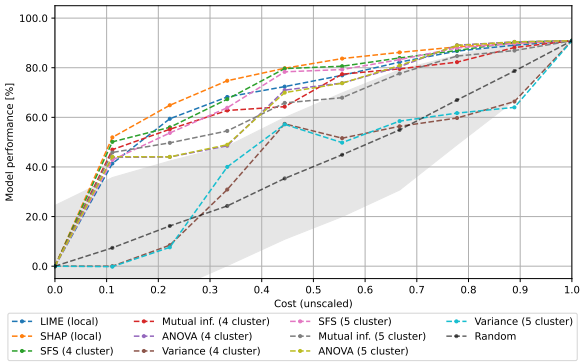


(d) Dynamic model-based

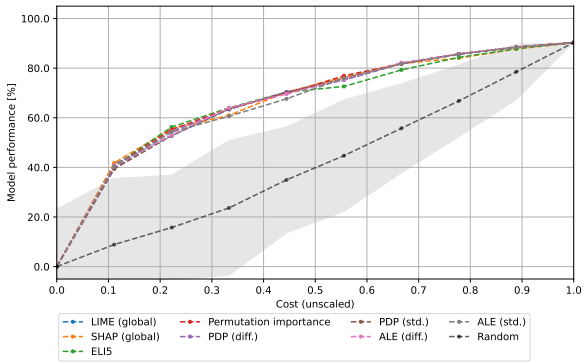
Figure 21: Results of cost-based evaluation on Gas (CO2) (Kaya et al. 2019) dataset.



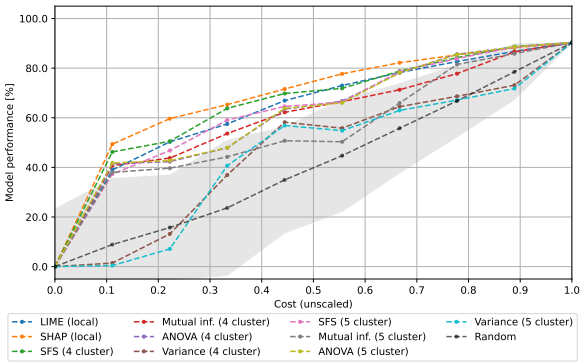
(a) Static data-based



(b) Dynamic data-based

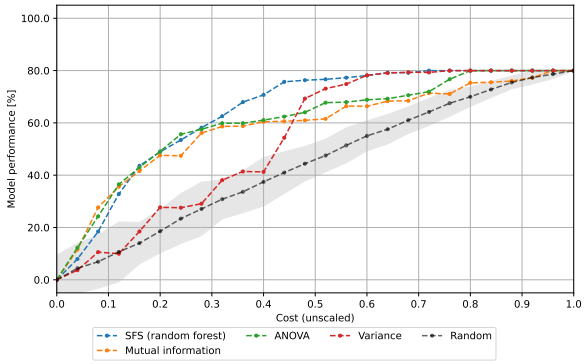


(c) Static model-based

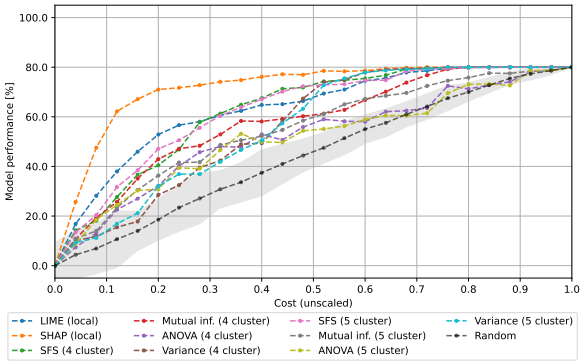


(d) Dynamic model-based

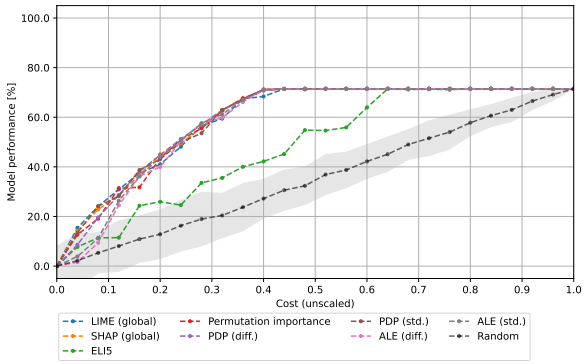
Figure 22: Results of cost-based evaluation on Gas (NOx) (Kaya et al. 2019) dataset.



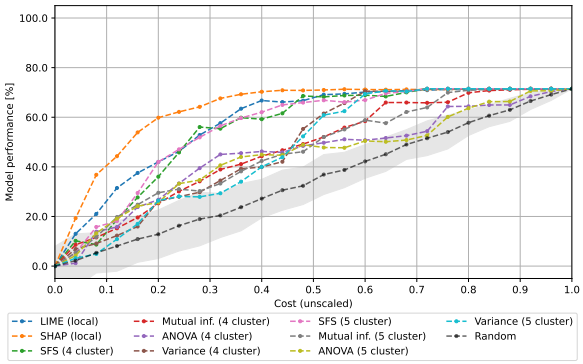
(a) Static data-based



(b) Dynamic data-based

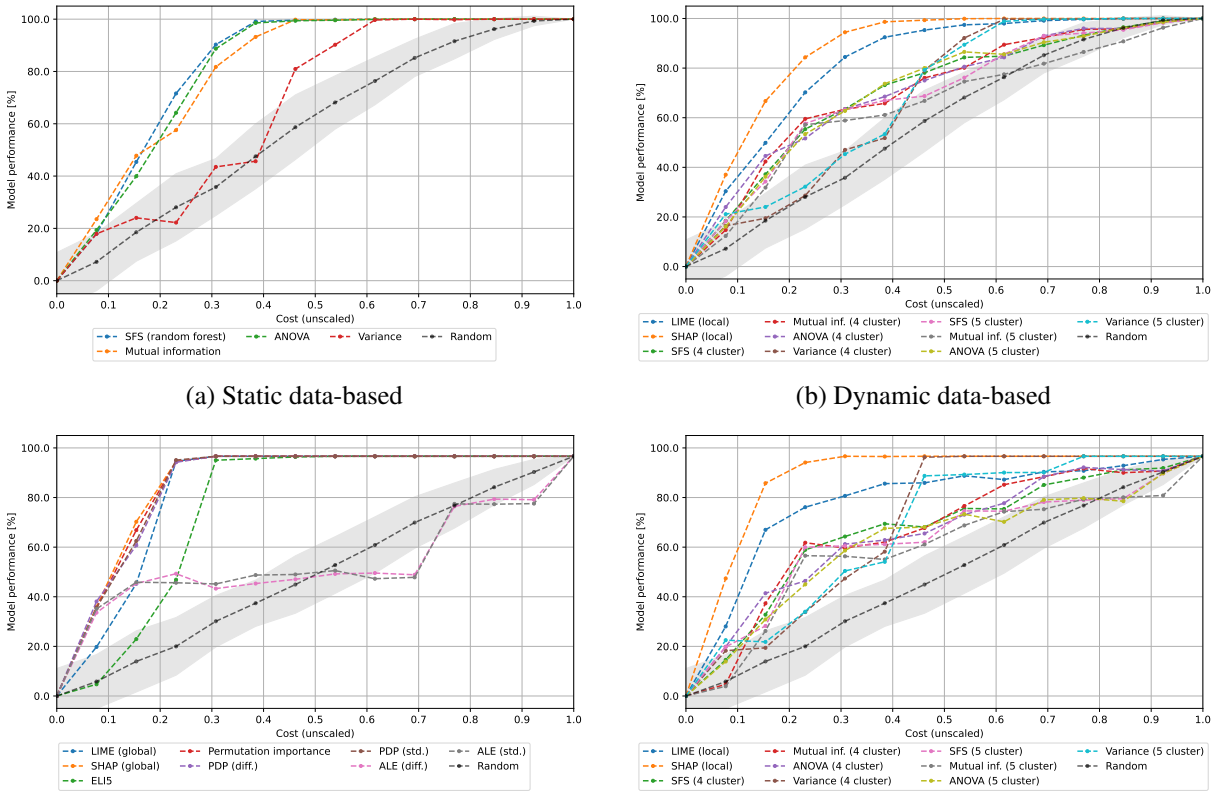


(c) Static model-based



(d) Dynamic model-based

Figure 23: Results of cost-based evaluation on Heart Disease (Janosi et al. 1988) dataset.



(a) Static data-based

(b) Dynamic data-based

(c) Static model-based

(d) Dynamic model-based

Figure 24: Results of cost-based evaluation on Wine (Aeberhard and Forina 1991) dataset.

  - *Context-Awareness:* This construct distinguishes between *static evaluations*, which assess data in isolation, and *dynamic evaluations*, which incorporate contextual factors like use cases or environmental conditions.

  These constructs are operationalized through seven design requirements, ensuring relevance across diverse data marketplace scenarios.

- **Principles of Form and Function:** The design incorporates seven design requirements and four tentative design principles. These principles guide the development of the DQSM and are instantiated through six design features, enabling an iterative and structured approach to artifact development. The evaluation strategy comprises five episodes and is outlined in Figure 1.

- **Justificatory Knowledge:** The DQSM design leverages justificatory knowledge derived from established fields:

  - *Feature Selection Techniques:* Methods from the field of machine learning that rank data features based on their relevance to predictive tasks.
  - *Explainable Artificial Intelligence (XAI):* Techniques that provide transparency and interpretability in decision-making.
  - *Active Feature Value Acquisition:* Approaches that prioritize features based on their utility for decision-making tasks.

  These knowledge bases ensure theoretical grounding when deriving design principles.

- **Testable Proposition:** To validate the DQSM, we formulated the following testable proposition:

  Implementing a DQSM that integrates feature selection and XAI techniques is technically feasible and enables accurate and context-sensitive quantification of data asset value.

This proposition was tested through evaluations on diverse datasets and expert interviews, demonstrating feasibility, consistent artifact performance, and practical utility.

- **Artifact Mutability:** Acknowledging the rapid advancements in machine learning and XAI, we designed the DQSM with mutability to accommodate evolving techniques and emerging marketplace requirements. This ensures adaptability and future-proofing of the artifact.
- **Principles of Implementation:** The DQSM is instantiated through six design features derived from the guiding principles. These features were iteratively refined across evaluation episodes, ensuring practical scalability and relevance to real-world use cases.
- **Expository Instantiation:** The artifact developed to operationalize the DQSM serves as a proof of concept, demonstrating its potential to address challenges in data asset valuation and selection. This instantiation supports data consumers in identifying valuable data assets while fostering trust and utility in data marketplaces.