To select or not to select? The role of meta-features selection in meta-learning tasks with tabular data

Irina Deeva[1 [0000−0001−8679−5868]] and Alena Kropacheva[1 [0009−0002−7703−4762]]

ITMO University, St. Petersburg, 197101, Russia

Abstract. In meta-learning tasks with tabular data, the choice of meta-features significantly impacts model performance and interpretability. This study investigates the necessity and methods of meta-feature selection in the context of meta-learning, particularly for tabular datasets. We address the fundamental question: Is it better to select a subset of meta-features or use the entire feature set? We examine various selection techniques, including filter, wrapper, and embedded methods, as well as a novel causal-based approach utilizing counterfactual reasoning. Our experiments demonstrate that feature selection generally enhances performance, with causal-based methods, especially those leveraging counterfactual generation, showing superior efficiency and generalizability. Furthermore, we explore how these methods fare under shifts in data, particularly when non-informative features are added. The results reveal that the counterfactual method maintains high efficacy across different meta-learners and exhibits a favorable balance between model performance and interpretability. These findings underscore the importance of meta-feature selection in improving the adaptability and transparency of meta-learners for tabular data tasks.
Code and supplementary materials for this research are available on GitHub: `https://github.com/ITMO-NSS-team/MetaSelect`.

Keywords: Meta-learning · Meta-features selection · Causal-based methods · Counterfactual reasoning · Tabular data.

1 Introduction

Motivation. In the world of machine learning (ML), there exist domains that are characterised by a substantial and varied array of models, such as machine learning on tabular data. This domain is replete with models of diverse classes, giving rise to a plethora of research endeavours aimed at ascertaining the relative merits of these models. A notable example of this is the investigation into whether deep models possess an advantage in the context of tabular data [5], [23], [24]. Recent studies have indicated that a more effective strategy might be to focus on optimizing the hyperparameters of a specific tabular model rather than on selecting the most optimal model [16]. This is understandable given the complexity and time-consuming nature of selecting the best model. However, there

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

exist methods and approaches, such as meta-learning, that aim to simplify this process. The fundamental components of meta-learning encompass the meta-characteristics of the data and the meta-learner itself, whose function is to facilitate the adaptation of performance information from disparate models to novel data without the necessity of directly training and evaluating those models [10]. The creation of an initial space of meta-characteristics (meta-features) is pivotal to the success of the meta-learning task [20]. Nevertheless, there is a paucity of research on the advisability of reducing the initial meta-characteristics space [18],[9], despite the plethora of studies on meta-characteristics selection [8], [19] [7].

Contribution. Whilst the reduction of the dimensionality of the meta-features space through selection appears reasonable in terms of efficiency and interpretability, a fundamental investigation of this problem was deemed necessary in order to answer the main research question: *when it comes to meta-features, is it better to pick and choose, or is it always better to train on the whole set of meta-features?* In contrast to previous studies, our analysis encompasses not only the impact of selection on meta-learner performance, but also the impact on interpretability. To this end, we have suggested metric for the latter and have analysed a large number of meta-features selection approaches. Our analysis enables us to draw conclusions about the effectiveness and generalisability of these approaches to different meta-learners. In addition, an approach for meta-features selection based on causal analysis and counterfactual reasoning is proposed. The experimental studies demonstrate the high efficiency and generalisability of this approach.

2 Related works

2.1 Meta-learning for tabular models

Given that the focus of this paper is on tabular data, it would be prudent to analyse studies that address the problem of selecting algorithms for such data. In essence, meta-learning based on meta-characteristics bears resemblance to a conventional machine learning task, with meta-features serving as predictors and the outcomes of machine learning models (e.g., performance) designated as the target variable. In [23], the authors conducted experiments on a benchmark of 111 datasets using 20 machine learning models for tabular data classification. The meta-model employed is a logistic classifier, the purpose of which is to minimise the error of predicting the win of a particular model based on the characteristics of the data. The authors [25] propose an alternative approach, which involves the initial reduction of the dimensionality of the original meta-features space. This is followed by the training of a meta-learner SVM in this space. In a separate publication [22], the authors conducted an investigation into a variety of models, encompassing both classical models (logistic regression, decision tree models) and deep models (FT-transformer, MLP) as meta-learners. The authors observe that while deep models demonstrate optimal efficiency, they exhibit a concomitant loss in interpretability. Also the authors [29] propose an alternative

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

predictive problem, namely the prediction of changes in the performance curve based on initial points. The proposed meta-model involves the identification of a mapping between meta-characteristics and initial values of dynamics, as well as validation statistics.

It is evident that the domain of meta-learning for tabular data is undergoing continuous development. Currently, a prevalent practice is the selection of either classical or deep models for tabular data, as the efficacy of these methods remains a subject of ongoing research. Meta-learning has emerged as a potential solution to automate this decision-making process. It is also noteworthy that in the vast majority of papers, authors select meta-features in one way or another, yet rarely elucidate the rationale behind their selection and the underlying principles that guided their decision-making. This underscores the necessity and pertinence of undertaking a more comprehensive investigation into the matter of reducing the set of meta-features as a part of meta-learning pipeline.

2.2 Meta-features selection methods

With regard to the selection of meta-features, it is first necessary to recognise that the task of meta-features selection is a particular instance of the more general task of feature selection in machine learning. Consequently, all the established approaches to feature selection are applicable to the selection of meta-features. The classification of feature selection methods can be divided into three distinct groups: $filter$, $embedded$ and $wrapper$. Each of these groups possesses its own advantages and disadvantages. Filter methods comprise a range of approaches, including correlation-based selection, selection based on statistical tests (e.g., the Chi-square test and the ANOVA F-test), and selection based on information criteria (e.g., mutual information) [11]. These methods are simple and computationally efficient, and can also be termed model-free. Wrapper methods address the feature selection problem through the lens of an optimisation problem. The quality of the machine learning model for which the features are selected is frequently employed as a metric to gauge the efficacy of the selection process. The optimisation algorithms employed in this context often encompass evolutionary algorithms [28], greedy algorithms [21], and more complex algorithms such as recursive feature elimination [27] or forward-backward selection with early dropping [4]. Wrapper methods have been shown to be efficient, but they are sensitive to the dimensionality of the source space and therefore not always computationally efficient. Finally, a group of embedded methods involves the direct selection of features during the training of machine learning (ML) models. The most popular approach here is lasso regression, and there are also its modifications, for example, LassoNet [12] and Deep Lasso [6]. The primary disadvantage of embedded methods is that they are model-free and thus contingent on the properties of the model itself.

In this paper, we propose the identification of an additional group of methods, which we designate as $causal - based$. These methods have application in both the filter and wrapper groups; however, we consider them as a discrete group, given that they are predicated on entirely distinct principles of causal machine

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

learning.Current methods for the selection of features do not invariably elucidate the features that engender a result in the data. In such instances, the principles of causal machine learning can prove beneficial. The majority of causal-based methods are predicated on the identification of the Markov Blanket (MB) of the target variable, since it is the attributes included in the MB that contain the main information about the target variable [31]. MB search-based approaches are informative, but they are also sensitive to a large number of features. Another group of methods is based on causal inference, for example, feature selection is based on calculated Average Treatment Effect values [15] or on counterfactual reasoning [30], [13]. To the best of our knowledge, causal approaches have not yet been investigated by authors for the task of meta-features selection, so in this study we are going to fill this gap by comparing them with other classical approaches to meta-features selection.

3 Backgrounds

The goal of meta-feature selection is to identify an optimal subset of features $S \subseteq F$ from a candidate set $F = \{X_1, X_2, \ldots, X_n\}$ that maximizes the predictive performance of a meta-learner M while minimizing subset size. This is formalized as a regularized optimization problem:

$$S^* = \arg\min_{S \subseteq F} \left[\mathcal{L}\big(M(S)\big) + \lambda |S| \right], \tag{1}$$

where $\mathcal{L}$ represents the meta-learner's loss function and λ controls the trade-off between model performance and meta-feature set cardinality. The mathematical formulations of the methods to be investigated in this paper are described below.

3.1 Filter Methods

Filter methods rank features using statistical measures of relevance:

– **Spearman's Rank Correlation:**

$$\rho(X_i, Y) = \left| \frac{\mathrm{Cov}(\mathrm{rank}(X_i), \mathrm{rank}(Y))}{\sigma_{\mathrm{rank}(X_i)} \sigma_{\mathrm{rank}(Y)}} \right| \tag{2}$$

– **Mutual Information (MI):**

$$\mathrm{MI}(X_i, Y) = \sum_{x \in X_i} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x, y)}{P(x) P(y)} \right) \tag{3}$$

– **ANOVA F-Test (continuous vs categorical):**

$$F(X_i, Y) = \frac{\frac{1}{K-1} \sum_{k=1}^{K} n_k (\bar{X}_i^{(k)} - \bar{X}_i)^2}{\frac{1}{N-K} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (X_{ij}^{(k)} - \bar{X}_i^{(k)})^2} \tag{4}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

3.2 Embedded Methods

Feature selection is integrated into the learning algorithm:

- **Lasso Regression (L1 Regularization):**

$$\beta^* = \arg\min_{\beta} \left\{ \frac{1}{2N} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{5}$$

  Non-zero coefficients ($\beta_j \neq 0$) indicate selected features.
- **XGBoost Feature Importance:**

$$\text{Importance}(X_j) = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_j(t)} \text{gain}(s), \tag{6}$$

  where $\mathcal{S}_j(t)$ denotes splits on feature X_j in tree t. Non-zero $Importance(X_j)$ indicates sekected features.

3.3 Wrapper Methods

Recursive Feature Elimination (RFE). RFE iteratively removes the least important features:

1. Initialize with full feature set $S_0 = F$
2. For each iteration t:
   - Train model M_t on current features S_t:

$$\theta_t = \arg\min_{\theta} \mathcal{L}(M_t(S_t; \theta)) \tag{7}$$

   - Compute feature importances $\{w_i^{(t)}\}$
   - Remove lowest-ranked features:

$$S_{t+1} = S_t \setminus \left\{ X_j \mid w_j^{(t)} = \min_{X_i \in S_t} w_i^{(t)} \right\} \tag{8}$$

3. Terminate when $|S_t| = k$ (desired subset size). Here desired subset size is equal to half of the original number of meta-features.

3.4 Causal-based meta-feature selection

Average treatment effect based selection. For robust estimation of heterogeneous treatment effects in meta-feature selection, we employ the CausalForestDML estimator [3], which combines double machine learning (DML) with causal forests [2]. For a meta-feature X_j acting as treatment variable T, and outcome Y representing algorithm performance, we model:

$$Y = \theta(X_j)T + g(X_{-j}) + \epsilon, \quad T = f(X_{-j}) + \eta, \tag{9}$$

where $X_{-j} = F \setminus \{X_j\}$ are confounders, $g(\cdot)$ and $f(\cdot)$ are nuisance functions, $\epsilon \perp \eta \perp X$.

The CausalForestDML estimator implements the following steps:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

1. **Orthogonalization**: Fit regularized models to estimate

$$\tilde{Y} = Y - \hat{g}(X_{-j}) \tag{10}$$

$$\tilde{T} = T - \hat{f}(X_{-j}) \tag{11}$$

2. **Causal Forest**: Estimate conditional average treatment effect (CATE) via

$$\hat{\theta}(x_j) = arg\min_{\theta} \sum_{i=1}^{N} \alpha_i(x_j)(\tilde{Y}_i - \theta\tilde{T}_i)^2, \tag{12}$$

where weights $\alpha_i(x_j)$ are determined by forest similarity kernel.

The feature importance for X_j is computed as:

$$\text{ATE-Importance}(X_j) = \mathbb{E}\left[\hat{\theta}(X_j)\right] \cdot \sqrt{\text{Var}(\hat{\theta}(X_j))}, \tag{13}$$

prioritizing features with both large average effects and effect heterogeneity.

Counterfactual based selection. For causal interpretation through intervention analysis, we propose a counterfactual feature selection mechanism. Existing counterfactual generation-based feature selection approaches typically utilise these counterfactuals as supplementary data [13]. In contrast, the present study proposes a direct utilisation of the counterfactual generation efficiency as a feature selection criterion. The rationale underlying this approach is that if this generation is successful in a given feature space, it can be assumed that these are the features that influence the target variable.

Given a classifier f and meta-feature $s_j \in S$, we formulate counterfactual generation as the optimization problem:

$$\delta^* = arg\min_{\delta} \underbrace{\mathcal{L}_{\text{pred}}\left(f(x+\delta), y^{\text{target}}\right)}_{\text{Prediction Loss}} + \lambda \underbrace{\frac{1}{N}\sum_{i=1}^{N} \frac{|x_i^{\text{cf}} - x_i^{\text{orig}}|}{\text{MAD}}}_{\text{Robust Distance Loss}} \tag{14}$$

where δ is the perturbation, y^{target} is the opposite to $f(x)$ label, λ controls minimal intervention strength, $\text{MAD} = \text{median}(|x_i - \text{median}(x)|)$ scales perturbations by feature robustness. It is asserted that a counterfactual generation which compelled the classifier to reverse the label, whilst concomitantly effecting only negligible alterations to the data, is to be regarded as effective. Then the feature selection criterion evaluates:

$$\text{Efficiency}(s_j) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[f(x_i) \neq f(x_i + \delta_i^*)\right] \tag{15}$$

A meta-feature s_j is selected if $\text{Efficiency}(s_j) > \epsilon$, where ϵ is a tolerance threshold. In this study, the linear classifier was utilised as the classification model (f), while the cross-entropy function was employed to calculate the prediction loss. Additionally, the Adam optimiser, with a learning rate set to 0.01, was employed for optimisation purposes.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

4 Design of Experiments

4.1 Dataset and meta-features description

We use experiment results provided in paper [16]. Authors present metrics for 19 algorithms (including GBDTs, neural networks and baselines) on 176 classification datasets from OpenML [26]. The results table is organized as follows: datasets are split into 10 folds, with each fold corresponding to a row containing metrics for every model with different hyperparameter sets. The evaluated metrics include logarithmic loss, AUC, accuracy, F1 score, and runtime (in seconds), calculated for three data splits: train, validation, and test. In our study, we focus on the F1 score from the test splits as the target metric for the following models: Logistic Regression, Random Forest, XGBoost, MLP, ResNet, and FT-Transformer.

Experiment results also include meta-features dataset. For each dataset fold meta-features are calculated with Python package PyMFE [1]. They include general features (such as number of attributes, number of distinct classes, etc), statistical features (such as skewness or kurtosis), information-theoretic features (such as noisiness of attributes or joint entropy), landmarking features (such as performance of the Naive Bayes classifier) and model-based features (such as number of leaf nodes in the Decision Tree model). Some meta-features are represented as vectors (e.g., maximum value from each attribute). These are aggregated using various functions, including average, maximum, minimum, standard deviation, kurtosis, skewness, and interquantile range.

The meta-dataset used in our research consists of meta-features and target metrics for each dataset. The data preprocessing steps include:

- Aggregating meta-features by folds using median values,
- Removing meta-features with more than half of the values undefined,
- Excluding meta-features with large absolute values,
- Eliminating constant meta-features,
- Filtering meta-features based on the aggregation function,
- Removing duplicate columns and rows,
- Excluding datasets with undefined target metrics,
- Scaling data using Yeo-Johnson transform,
- Filtering meta-features with large VIF [17].

The final meta-dataset consists of 134 datasets, 123 meta-features.

Meta-learning tasks and meta-learners. In the present study, the following meta-learning task was considered: predicting the F1 score of a ML model. The task was formulated as binary classification task. The F1 score was predicted as follows: if the F1 score exceeded the median of the target scores then 1 was assigned; otherwise, 0 was assigned. The models employed as meta-learning models were KNN Classifier (6 neighbors, uniform weight function, leaf size: 40), XGBoost Classifier (maximal depth: 7, learning rate: 0.1, 50 estimators, evaluation

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

metric: accuracy) and MLP Classifier (hidden layer size: 25, logistic activation function, L-BFGS solver, strength of the L2 regularization: 0.05, adaptive learning rate with initial value 0.05).

5 Results

The following research questions are formulated in this study:

- RQ1. Should we make a selection of meta-features, and how do the different ways of selecting features compare?
- RQ2. How do methods behave under shifts in the data when non-informative features are added?
- RQ3. How do the methods relate to each other in terms of the interpretability of the selected features?

5.1 Study of the performance of feature selection methods

Initially, a baseline study was conducted in order to ascertain the performance of the selection methods. These were then compared to their profit performance on all meta-features. The results are presented in figure 1. The initial conclusion that can be drawn is that feature selection almost invariably improves the quality of the result, thereby demonstrating its necessity. With regard to the performance of feature selection methods, causal-based approaches, and in particular the proposed algorithm based on **counterfactual reasoning**, demonstrate favourable results. Indeed, the algorithm is among the top three methods in more than half of the results. In the context of a group of methods, filter-based and causal-based methods demonstrate optimal performance, with wrapper approaches and embedded methods exhibiting average performance.

A thorough analysis of the performance of the methods with respect to the target type reveals that the method based on counterfactual generations exhibited equivalent proficiency in predicting the productivity of classical models and the performance of deep models. With regard to meta-learning models, the counterfactual method demonstrated notable efficacy in the MLP and XGBoost models.

A study was also conducted on the computational complexity of feature selection methods. As illustrated in the figure 2, the outcomes of measuring the execution time of selection methods in relation to the dimensionality of the initial feature space are presented. As would be anticipated, causal-based methods demonstrate the greatest longevity and exhibit increased sensitivity to the augmentation in the number of features.

5.2 Feature Selection Under Data Shifting Conditions

In order to undertake a comprehensive investigation into the performance of feature selection methods, experiments were conducted with the objective of understanding how the methods respond to shifts in the data, namely the addition

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

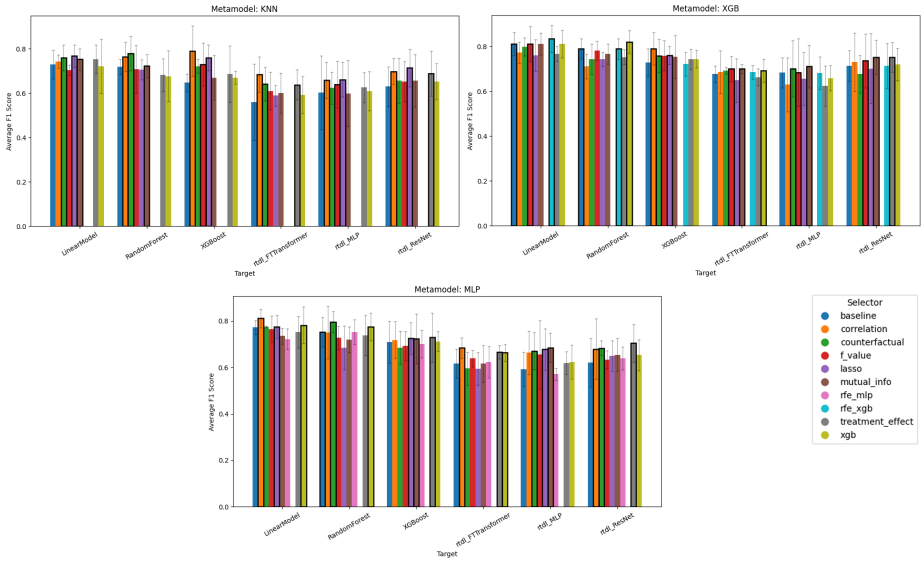

Fig. 1. Results comparing feature selection methods, here *base* is a run on all meta-features. In each category, the top 3 methods are highlighted with a black box.

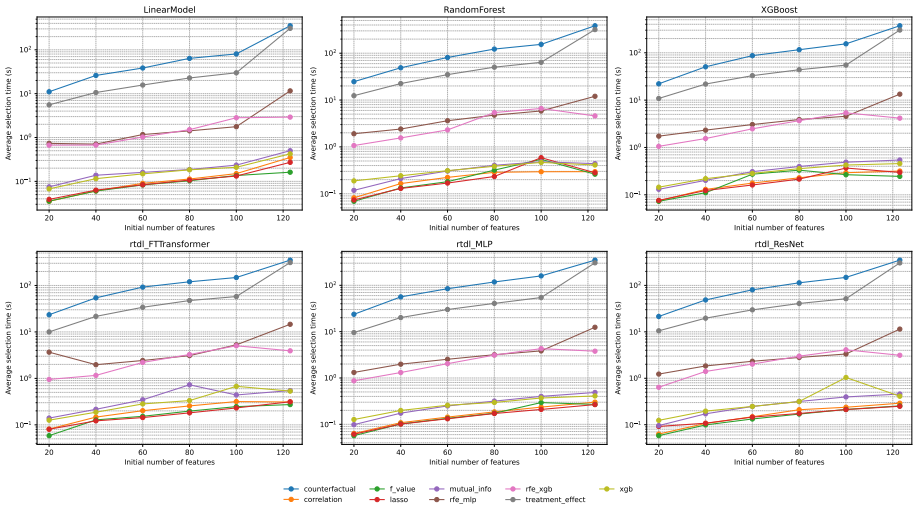

Fig. 2. Results comparing the running time of feature selection methods as a function of the initial number of features.

of uninformative features. In order to achieve this, the logic of the experiments presented in the paper [6] was utilised. Three categories of uninformative features were generated based on the said study:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

- Random features - generated from a random Gaussian distribution.
- Corrupted features - the original features are randomly selected and Gaussian noise is added to them.
- Second-order features - initial features are randomly selected and new features generated from them by means of squaring, addition, multiplication, etc.

In the course of the experiment, 50% of uninformative features were appended to the dataset, with the objective of investigating the performance of meta-features selection methods in terms of classification performance. The results of the experiments are presented in figure 3. In order to facilitate the perception of results, the mean value was calculated for each target type. This enabled the prediction of the performance of both the classical and deep models. It is evident that the quality remains largely consistent with the calculated efficiency outlined in the preceding section (section 5.1). The f-value statistical test emerged as the most robust approach when confronted with variations in uninformed features. Nonetheless, the counterfactual generation method consistently yielded commendable results, consistently ranking within the top three in over half of the cases examined. The performance of the embedded methods (lasso and xgb) was found to be suboptimal in the addition of uninformative features.

5.3 Feature Selection and Interpretability

In addition, an investigation was conducted into feature selection methods with regard to the interpretability of the results obtained. It is intuitive to assume that if features are selected, their influence can be more easily analysed at a later stage. In order to evaluate the interpretability of the results obtained after feature selection, a metric **importance fraction score** is introduced (Eq. 16). The prevailing logic in this context posited that an elevated concentration of importance among a reduced number of meta-features would result in enhanced interpretability. Consequently, the significance of each feature (I) was calculated using the SHAP method [14]. The ratio of the sum of the top five features significances to the sum of the significances of all the selected features was then determined. Consequently, if the significance of the features was "distributed" over all the selected features, this indicated a low interpretability of the selected features and resulted in low importance fraction scoring.

$$Importance\ fraction\ score = \frac{\sum_{i=1}^{k} I(s_i)}{\sum_{i=1}^{n} I(s_i)} \tag{16}$$

The results of the feature importance score comparison are displayed in the figure 4. It is evident that the lasso and xgboost methods are the most prominent. However, this can be readily explained by the fact that these methods typically select a minimal number of meta-features (no more than 10), resulting in a high importance fraction score. It is noteworthy that the counterfactual and correlation methods typically select a comparable number of features (figure 5),

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

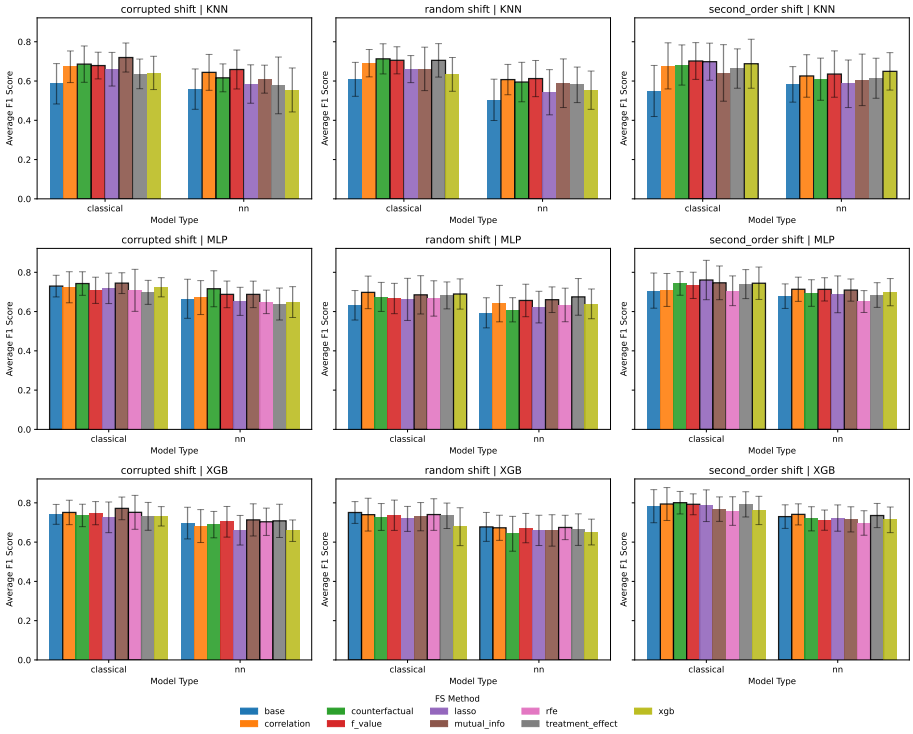

Fig. 3. Results comparing feature selection methods with different types of uninformative features, here *base* is a run on all meta-features. In each category, the top 3 methods are highlighted with a black box.

yet the importance fraction score of the counterfactual method is marginally higher on average. This finding suggests that the counterfactual method does indeed select causal features. The absence of feature selection naturally engenders low interpretability values of the results, which once again confirms the need to select meta-features.

6 Conclusion and Discussion

In this study, we examined the impact of meta-feature selection on the performance and interpretability of meta-learning models for tabular data. Our investigation addressed the fundamental question of whether it is more advantageous to select a subset of meta-features rather than relying on the entire feature set. Through extensive experimentation across various selection methodologies—including filter, wrapper, embedded, and our proposed causal-based counterfactual approach—we demonstrated that judicious feature selection can substantially enhance both predictive performance and the clarity of model interpretations.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

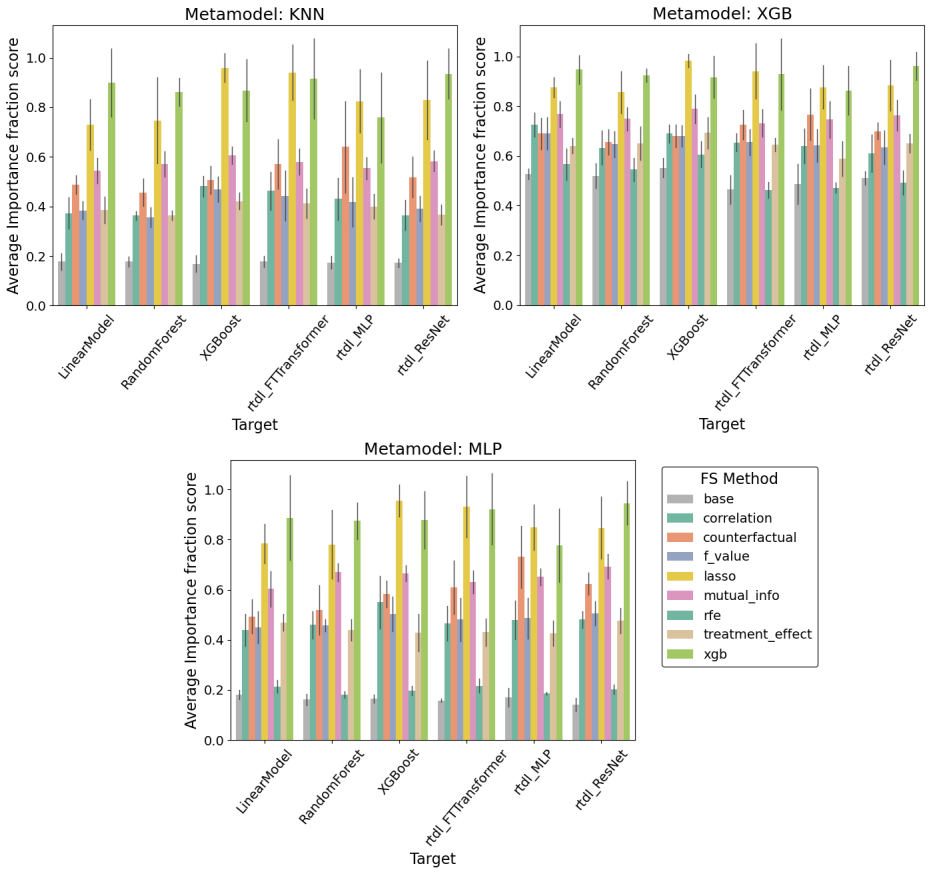

Fig. 4. Results of comparison of importance fraction score for different feature selection methods, here *base* is a run on all meta-features.

Our experimental results consistently indicated that meta-feature selection improves model outcomes. Notably, the causal-based methods, particularly the counterfactual generation approach, emerged as a robust solution, often ranking among the top performers across multiple meta-learners. This method not only maintained high efficiency under standard conditions but also exhibited strong resilience when datasets were augmented with uninformative features. Such robustness underscores its potential for real-world applications where data distributions may shift or noise is prevalent.

From an interpretability standpoint, our findings reveal that selecting a smaller, more informative set of meta-features allows for a concentrated distribution of feature importance. Methods such as Lasso and XGBoost tended to select a minimal number of features, thereby yielding high importance fraction scores. However, the counterfactual method, while selecting a comparable number of features to correlation-based approaches, provided marginally higher

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

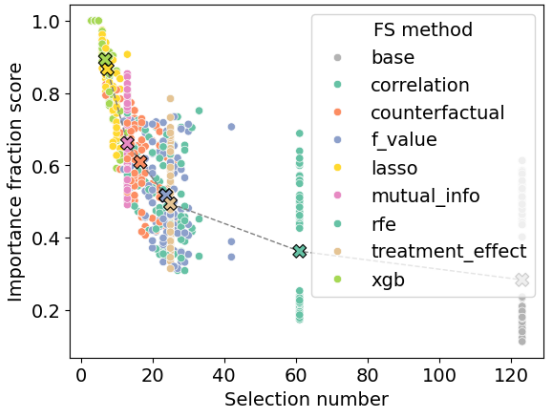

Fig. 5. Dependence of average importance fraction score on the number of selected meta-features, here the cross - feature selection method's cluster centroid.

interpretability. This suggests that causal-based selection not only filters out redundant information but also preserves the intrinsic causal relationships that are critical for understanding model behavior.

While the advantages of meta-feature selection are evident, our study also highlights several challenges. The performance of embedded methods was notably affected by the introduction of non-informative features, indicating a potential sensitivity to data quality. Additionally, the computational complexity of some selection methods, particularly those based on causal inference and counterfactual reasoning, warrants further investigation to ensure scalability and efficiency in larger datasets.

In conclusion, our work reinforces the importance of incorporating meta-feature selection into meta-learning pipelines, particularly for tasks involving tabular data. By carefully selecting meta-features, practitioners can achieve a more efficient and interpretable modeling process without sacrificing predictive accuracy. Future research may focus on refining causal-based selection techniques, exploring their applicability across diverse data modalities, and developing more computationally efficient algorithms. Ultimately, these efforts will contribute to the broader goal of automating and enhancing the decision-making process in machine learning model selection.

7 Acknowledgments

This research is financially supported by The Russian Scientific Foundation, Agreement № 24-71-10093, `https://rscf.ru/en/project/24-71-10093/`.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

References

1. Alcobaça, E., Siqueira, F., Garcia, L.P., Rivolli, A., Oliva, J., de Carvalho, A.: MFE: Towards reproducible meta-feature extraction. Journal of Machine Learning Research **21**, 1–5 (Jan 2020)
2. Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. Annals of Statistics (2019)
3. Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., Syrgkanis, V.: Econml: A python package for ml-based heterogeneous treatment effects estimation (2021), `https://github.com/microsoft/EconML`
4. Borboudakis, G., Tsamardinos, I.: Forward-backward selection with early dropping. Journal of Machine Learning Research **20**(8), 1–39 (2019), `http://jmlr.org/papers/v20/17-334.html`
5. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE transactions on neural networks and learning systems (2022)
6. Cherepanova, V., Levin, R., Somepalli, G., Geiping, J., Bruss, C.B., Wilson, A.G., Goldstein, T., Goldblum, M.: A performance-driven benchmark for feature selection in tabular deep learning. Advances in Neural Information Processing Systems **36** (2024)
7. Filchenkov, A., Pendryak, A.: Datasets meta-feature description for recommending feature selection algorithm. In: 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT). pp. 11–18. IEEE (2015)
8. Garouani, M., Ahmad, A., Bouneffa, M.M.: Explaining meta-features importance in meta-learning through shapley values. In: 25th International Conference on Enterprise Information Systems (ICEIS 2023). vol. 1, pp. 591–598. SCITEPRESS-Science and Technology Publications (2023)
9. Kalousis, A., Hilario, M.: Feature selection for meta-learning. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 222–233. Springer (2001)
10. Khan, I., Zhang, X., Rehman, M., Ali, R.: A literature survey and empirical study of meta-learning for classifier selection. IEEE Access **8**, 10262–10281 (2020)
11. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A.: A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM transactions on computational biology and bioinformatics **9**(4), 1106–1119 (2012)
12. Lemhadri, I., Ruan, F., Abraham, L., Tibshirani, R.: Lassonet: A neural network with feature sparsity. Journal of Machine Learning Research **22**(127), 1–29 (2021)
13. Liu, D., Yue, X.: Efficient feature selection algorithm based on counterfactuals. In: 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE). pp. 541–544. IEEE (2024)
14. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`
15. Mangal, A., Rathore, S.S.: Ate-fs: An average treatment effect-based feature selection technique for software fault prediction. ACM Transactions on Intelligent Systems and Technology (2025)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21

16. McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., White, C.: When do neural nets outperform boosted trees on tabular data? Advances in Neural Information Processing Systems **36** (2024)
17. O'brien, R.M.: A caution regarding rules of thumb for variance inflation factors. Quality & quantity **41**, 673–690 (2007)
18. Pereira, G.T., Santos, M.R.d., de Carvalho, A.C.P.d.L.F.: Evaluating meta-feature selection for the algorithm recommendation problem. arXiv preprint arXiv:2106.03954 (2021)
19. Pinto, F., Soares, C., Mendes-Moreira, J.: Towards automatic generation of metafeatures. In: Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I 20. pp. 215–226. Springer (2016)
20. Rivolli, A., Garcia, L.P., Soares, C., Vanschoren, J., de Carvalho, A.C.: Meta-features for meta-learning. Knowledge-Based Systems **240**, 108101 (2022)
21. Sağbaş, E.A.: A novel two-stage wrapper feature selection approach based on greedy search for text sentiment classification. Neurocomputing **590**, 127729 (2024)
22. Shao, X., Wang, H., Zhu, X., Xiong, F., Mu, T., Zhang, Y.: Effect: Explainable framework for meta-learning in automatic classification algorithm selection. Information Sciences **622**, 211–234 (2023)
23. Shmuel, A., Glickman, O., Lazebnik, T.: A comprehensive benchmark of machine and deep learning across diverse tabular datasets. arXiv preprint arXiv:2408.14817 (2024)
24. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022)
25. Smith-Miles, K., Muñoz, M.A.: Instance space analysis for algorithm testing: Methodology and software tools. ACM Computing Surveys **55**(12), 1–31 (2023)
26. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: networked science in machine learning. ACM SIGKDD Explorations Newsletter **15**(2), 49–60 (Jun 2014)
27. Xia, S., Yang, Y.: A model-free feature selection technique of feature screening and random forest-based recursive feature elimination. International Journal of Intelligent Systems **2023**(1), 2400194 (2023)
28. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on evolutionary computation **20**(4), 606–626 (2015)
29. Ye, H.J., Liu, S.Y., Cai, H.R., Zhou, Q.L., Zhan, D.C.: A closer look at deep learning on tabular data. arXiv preprint arXiv:2407.00956 (2024)
30. You, D., Niu, S., Dong, S., Yan, H., Chen, Z., Wu, D., Shen, L., Wu, X.: Counterfactual explanation generation with minimal feature boundary. Information Sciences **625**, 342–366 (2023)
31. Yu, K., Guo, X., Liu, L., Li, J., Wang, H., Ling, Z., Wu, X.: Causality-based feature selection: Methods and evaluations. ACM Computing Surveys (CSUR) **53**(5), 1–36 (2020)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97629-2_21