Regression-Stratified Sampling for Optimized Algorithm Selection in Time-Constrained Tabular AutoML

Mehdi Bahrami, So Hasegawa, Lei Liu , Wei-Peng Chen
Fujitsu Research of America

FUJITSU

SPIGM

ICML International Conference On Machine Learning

TL;DR: a Regression-Stratified Sampling method with a PDF Energy metric for selecting optimized ML algorithms in Tabular AutoML

Introduction

ML algorithm is indispensable for tabular AutoML training. It can be expensive for large tabular datasets, especially under time constraints. One of the popular approaches for exploring ML algorithms is a simple random sampling approach. However, this approach can result in poor algorithm selection [1]. Let M be a Bayesian model in a supervised setting for the given input X to predict Y with a parameter of θ with $\mathbb{D}_y$ distribution as follows.

$$\mathcal{P}(y, \theta | x) = \mathcal{P}(y | x, \theta)\mathcal{P}(\theta)$$

Algorithm Selection

Our hypothesis in this study to be tested is:

$$\mathcal{PDF}(f(\mathcal{X}^{\rho})) \approx PDF(f(\mathcal{X})) \text{ if } PDF(\mathcal{Y}^{\rho}) \approx \mathcal{PDF}(\mathcal{Y})$$

$$\mathcal{M}^{o} = \underset{i=[1,..,n]}{\operatorname{argmin}}(\mathcal{A}(\mathcal{L}_i(\mathcal{D})))$$

$$\mathcal{M}^{\rho} = \underset{i=[1,..,n]}{\operatorname{argmin}}(\mathcal{A}(\mathcal{L}_i(\mathcal{D}^{\rho})))$$

$$\mathcal{L} = \mathcal{PDF}(f(\mathcal{X}^{\rho})) - PDF(f(\mathcal{X}))$$

PDF Energy Metric

$$\mathbb{S}(\hat{y}_i) = \begin{cases} \mathbb{D}(y_i) & \beta_i \le \hat{y}_i < \beta_{i+1} \\ -\mathbb{D}(y_i) * ||\beta_i - \hat{\beta}_i|| & \hat{y}_i < \beta_i \text{ or } \hat{y}_i > \beta_{i+1} \end{cases}$$

$$\mathbb{E}_X(\hat{y}) = \sum_{k=1}^{||\mathbb{D}^{\rho}||} \mathbb{S}(\hat{y}_k)$$

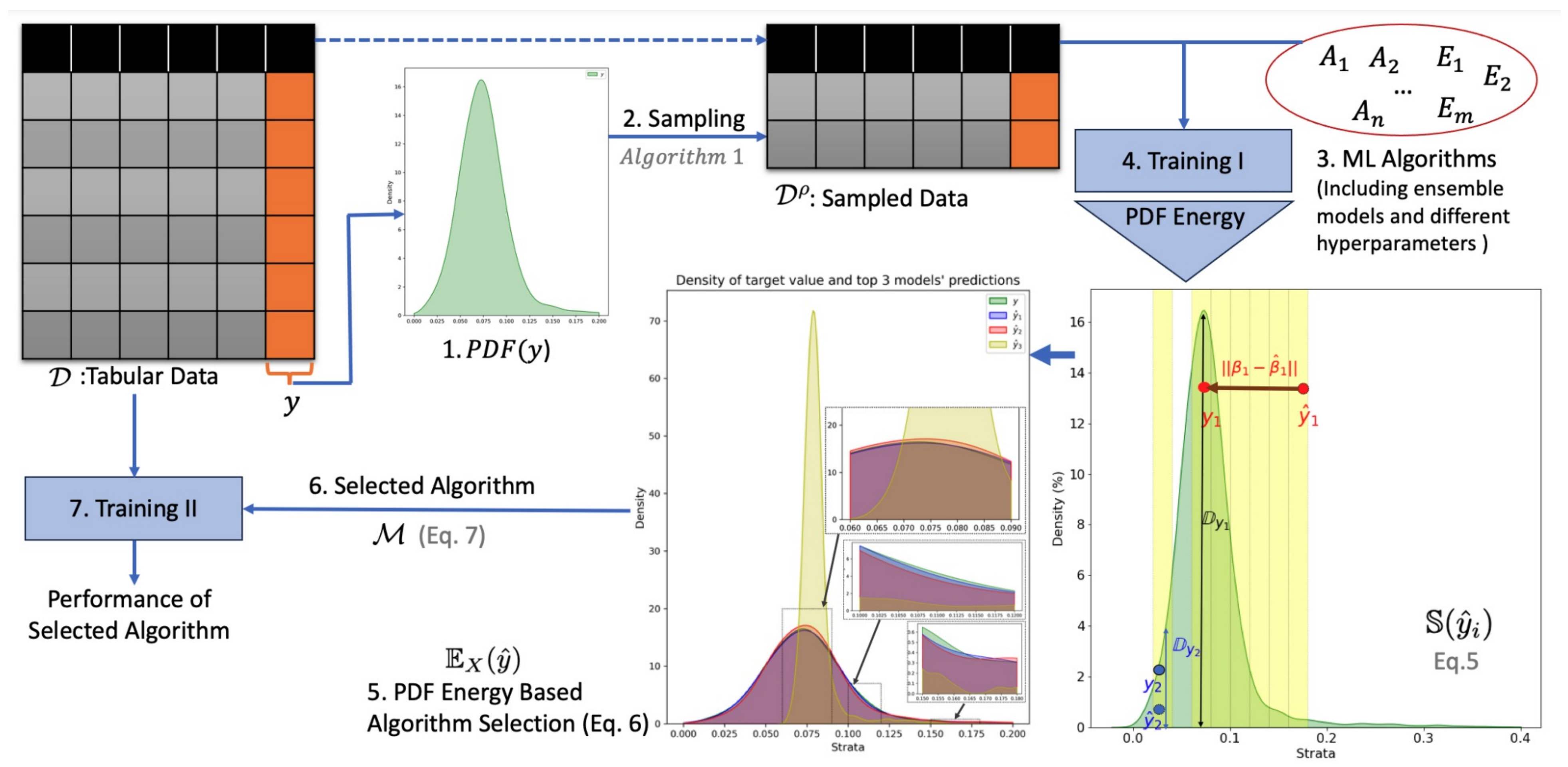
$$\mathcal{M} = \underset{\gamma \in \Gamma}{\operatorname{argmax}}(\mathbb{E}_X^{\gamma}(\hat{y}))$$

Experimental Setup

We utilized a tabular AutoML benchmark [2] and defined two sets of sub-benchmarks: #1 consists of 31 datasets, and #2 includes 14 real-world datasets for regression tasks.

AutoML	# of Choices
MLJAR	10
AutoGuon	6
H2O	5 categories includes 14 algorithms
Auto-Scikit Learn	15
TPOT	6
FLAML	6 (w/ hyperparameters)
Baseline+RSS (Our)	14

Regression Stratified Sampling



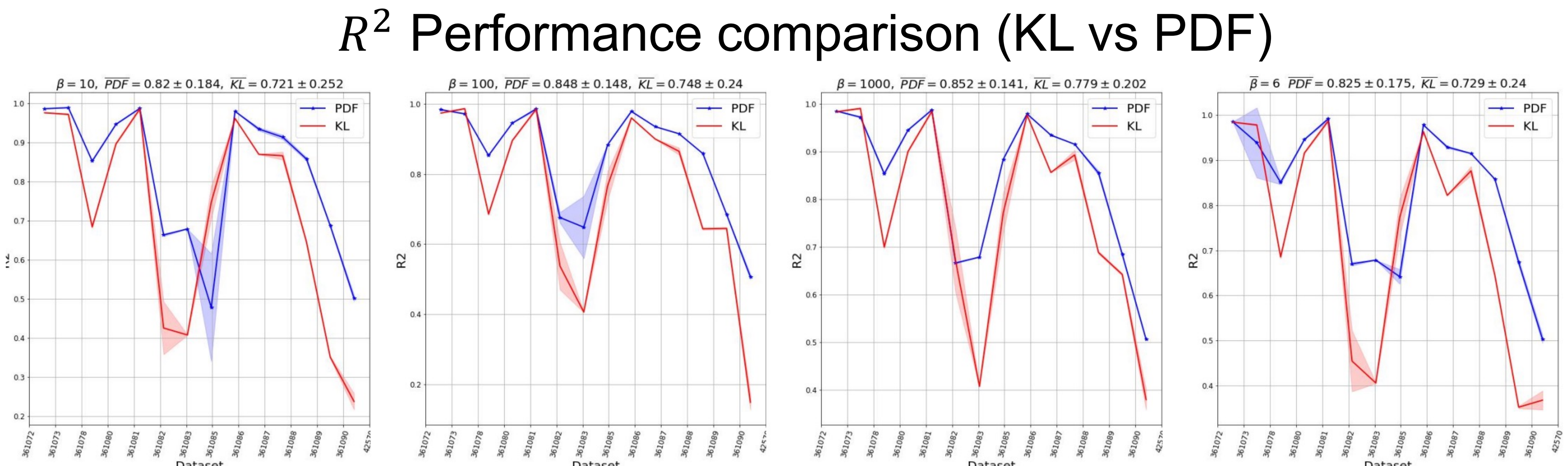
Overview of the proposed approach for algorithm selection

Experiment Results

Performance Comparison Simple Random Sampling vs RSS

			Final Evaluation on 25% hold-out data							
β			10		100		1000		Dynamic Stratified	
	Sampling Method	Eval. Method	$R^2\uparrow$	RMSE↓	$R^2\uparrow$	RMSE↓	$R^2\uparrow$	RMSE↓	$R^2\uparrow$	RMSE↓
Average	Random Sampling	Metric	0.8425 ± 0.012	9.6812 ± 0.359	0.8425 ± 0.012	9.6812 ± 0.3590	0.8425 ± 0.012	9.6812 ± 0.359	0.8425 ± 0.012	9.6812 ± 0.359
	PDF Sampling (our)	PDF Energy (our)	0.8183 ± 0.024	9.5147 ± 0.356	$\mathbf{0.8455 \pm 0.016}$	$\mathbf{9.5432 \pm 0.363}$	$\mathbf{0.8468 \pm 0.01}$	$\mathbf{9.6453 \pm 0.332}$	0.8254 ± 0.036	9.6757 ± 0.87
Total Number of Champions (Top rank across 14 possible datasets)		Metric	3	3	4	4	5	5	5	5
		Equal Results	2	2	0	0	1	1	1	1
		PDF Energy (our)	**9**	**9**	**10**	**10**	**8**	**8**	**8**	**8**

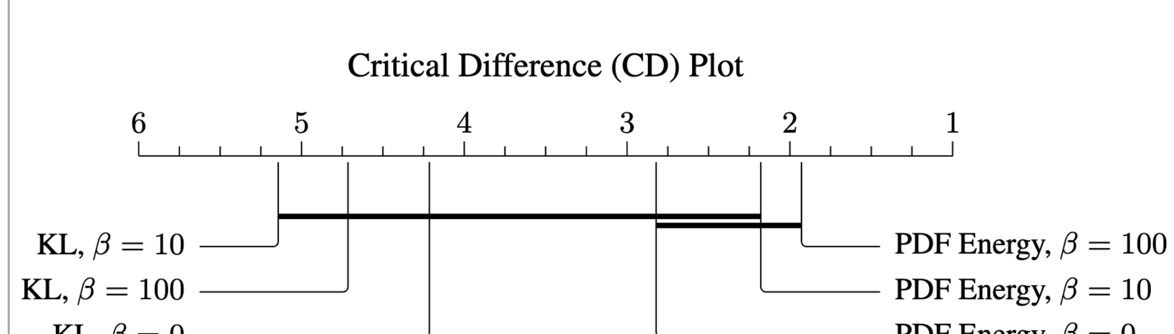
R^2 Performance comparison (KL vs PDF)



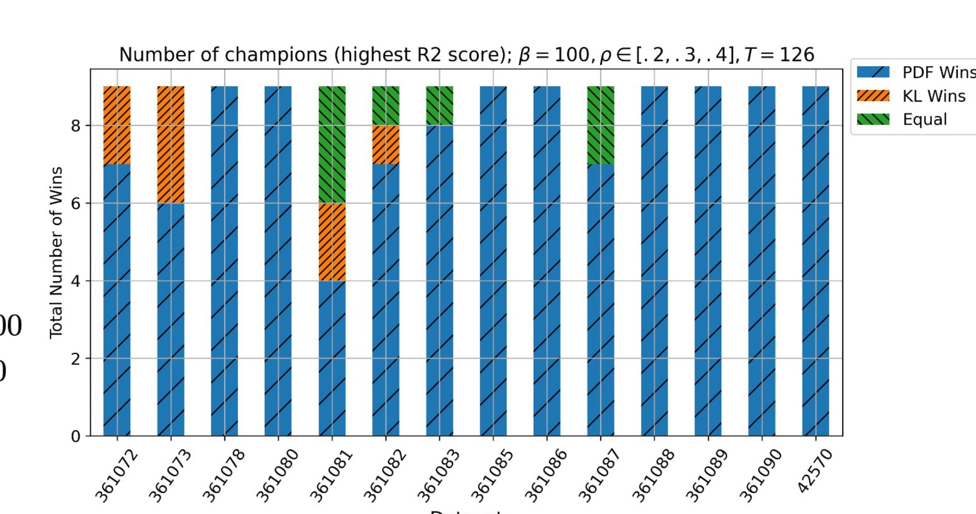
R^2 score of AutoML evaluation across 31 datasets

Time (s)	Baseline	MLJAR	FLAML	AutoSKLearn	H2O	TPOT	AutoGluon	RSS (our)
30	0.7601 ± 0.28	0.7662 ± 0.28	0.8069 ± 0.21	0.6607 ± 0.35	0.7885 ± 0.22	0.6597 ± 0.33	0.7047 ± 0.32	$\mathbf{0.8222 \pm 0.21}$
60	0.7663 ± 0.27	0.7764 ± 0.27	0.8152 ± 0.2	0.7179 ± 0.3	0.8004 ± 0.21	0.689 ± 0.32	0.7341 ± 0.3	$\mathbf{0.8239 \pm 0.2}$
120	0.7629 ± 0.28	0.7691 ± 0.28	0.8177 ± 0.2	0.7518 ± 0.27	0.8039 ± 0.22	0.7506 ± 0.27	0.7751 ± 0.27	$\mathbf{0.8242 \pm 0.2}$
180	0.7598 ± 0.28	0.7365 ± 0.28	0.819 ± 0.2	0.7819 ± 0.24	0.8054 ± 0.22	0.7618 ± 0.26	0.777 ± 0.27	$\mathbf{0.8243 \pm 0.2}$
300	0.761 ± 0.28	0.7277 ± 0.28	0.8217 ± 0.2	0.7923 ± 0.23	0.8131 ± 0.21	0.7748 ± 0.25	0.7716 ± 0.28	$\mathbf{0.8262 \pm 0.2}$

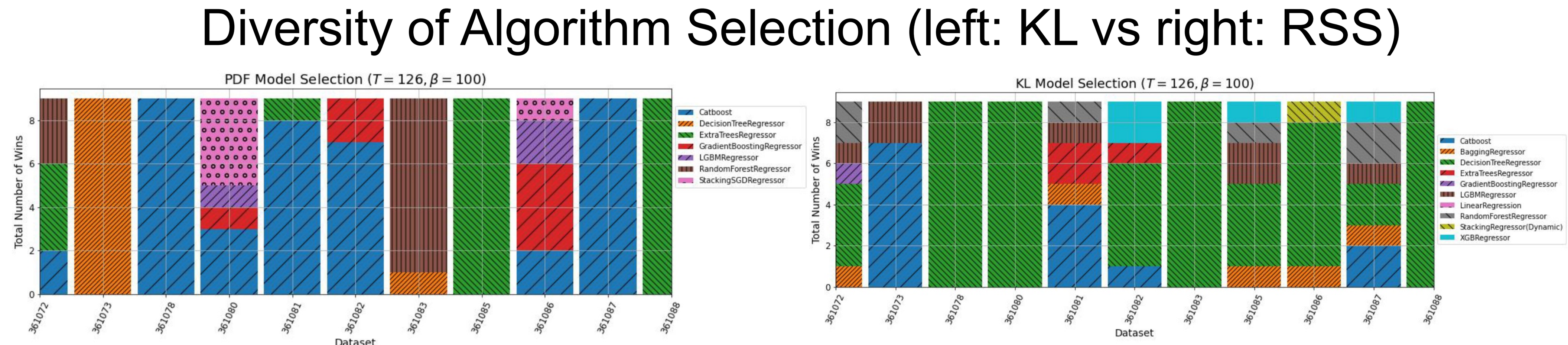
R^2 of algorithm selection (KL vs PDF Energy)



Algorithm selection (KL vs PDF)



Diversity of Algorithm Selection (left: KL vs right: RSS)



More details?
Paper
GitHub

Conclusion

Utilizing PDF in tabular AutoML for optimized algorithm selection is beneficial.

[1] Yakovlev, A., et al. Oracle automl: a fast and predictive automl pipeline. Proceedings of the VLDB Endowment, 13(12):3166–3180, 2020.
[2] Grinsztajn, Léo, E. Oyallon, and G. Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." *Advances in neural information processing systems* 35 (2022): 507-520.