Tunowalność Algorytmów ML

Automatyczne Uczenie Maszynowe – Projekt 1

Igor Kołodziej
Hubert Kowalski
Julia Przybytniowska

Spis treści

01

Cel projektu

Cel

- Optymalne hiperparametry **domyślne** - lepsze niż wartości *default* w *scikit-learn*
- 2 techniki losowania punktów:
  - Rozkład jednostajny
  - Metoda **bayesowska**
- **Tunowalność** algorytmów ML na 4 zbiorach danych [1]

[1] Zgodnie z definicjami z artykułu: *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*

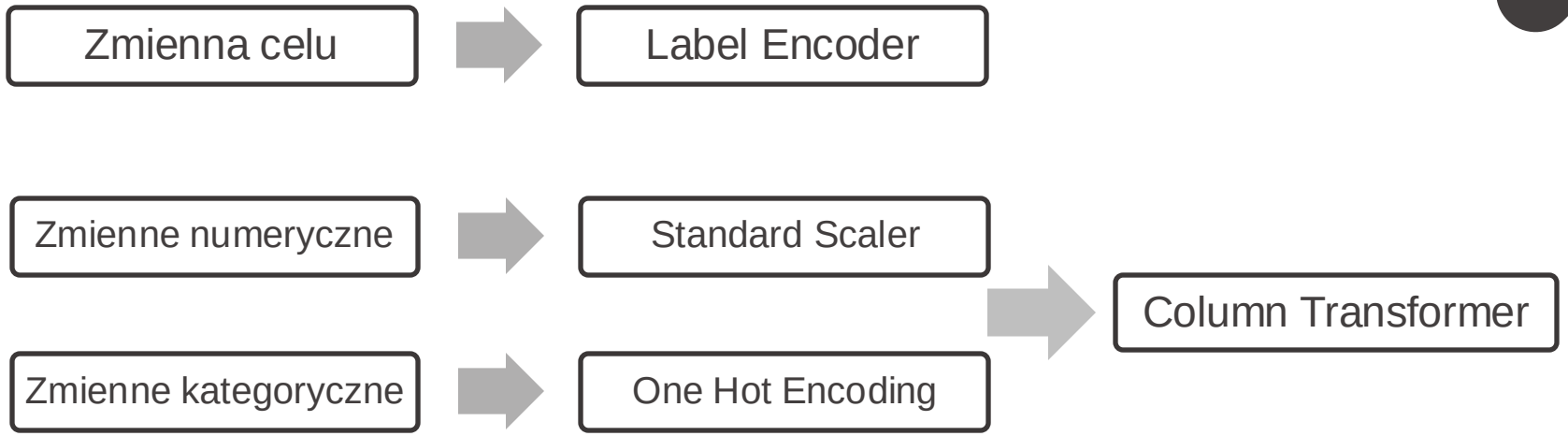
02

Zbiory Danych

Wybrane **zbiory danych**

- **Credit-g** (ID 31)
  1000 obserwacji, 20 zmiennych objaśniających, 0 wartości brakujących
- **Diabetes** (ID 37)
  768 obserwacji, 8 zmiennych objaśniających, 0 wartości brakujących
- **Wine** (ID 43980)
  2554 obserwacji, 11 zmiennych objaśniających, 0 wartości brakujących
- **Shrutime** (ID 45062)
  10000 obserwacji, 10 zmiennych objaśniających, 0 wartości brakujących

Preprocessing danych



Ewaluacja modeli

- Metryka

  Testowane: F1 Score, AUC, Accuracy → Wybrane: **AUC**

- Problem przeuczenia modelu
  - **CV**
  - **5 - fold**

- Powtarzalność wyników:
  - **RANDOM_SEEDS**

03

Modele i ich siatki hiperparametrów

Logistic Regression (ElasticNet)

Algorytm	Hyperparametr	Dolny zakres	Górny zakres
ElasticNet	C	1e-4	1e4
	penalty	elasticnet	
	l1_ratio	1e-4	1.0
	class_weight	balanced	
	max_iter	1500	
	solver	saga	

Tabela 1: Siatka hiperparametrów dla ElasticNet.

Extra Trees Classifier

Algorytm	Hiperparametr	Dolny zakres	Górny zakres
Extra Trees	n_estimators	10	1000
	criterion	gini, entropy, log_loss	
	bootstrap	True	
	max_samples	0.5	1.0
	max_features	0.1	0.9
	min_samples_leaf	0.05	0.25

Tabela 2: Siatka hiperparametrów dla Extra Trees Classifier.

Classifier

Algorytm	Hiperparametr	Dolny zakres	Górny zakres
XGBoost	n_estimators	10	2000
	learning_rate	1e-4	0.4
	subsample	0.25	1.0
	booster	gbtree	
	max_depth	1	15
	min_child_weight	1	128
	colsample_bytree	0.2	1.0
	colsample_bylevel	0.2	1.0
	reg_alpha	1e-4	512.0
	reg_lambda	1e-3	1e3

Tabela 3: Siatka hiperparametrów dla XGBoost Classifier.

04

Techniki losowania punktów

Metody losowania punktów

Technika losowania	Liczba iteracji
Random Search	300
Tree-Structured Parzen Estimator (Bayes Search)	(punkty startowe) 3 * 100 = 300

Biblioteka implementująca

OPTUNA

05

Wyniki eksperymentów

Optymalne wartości domyślne hiperparamerów - porównanie dla Logistic Regression

Params	C	l1_ratio	penalty	mean_AUC
Package defaults	1.0	-	l2	0.8155
Our best params	0.1869	0.00049	elasticnet	0.8160

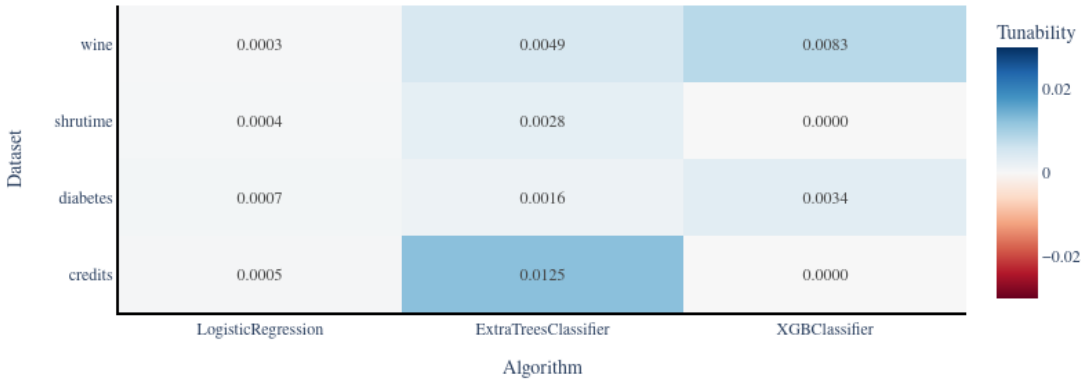
Wyniki optymalizacji modeli – ile możemy zyskać?

Model	Random Search mean best AUC	Bayesian Search mean best AUC	Default
LogisticRegression	0.8165	0.8166	0.8159
ExtraTreesClassifier	0.7979	0.8047	0.7946
XGBClassifier	0.8433	0.8527	0.8403

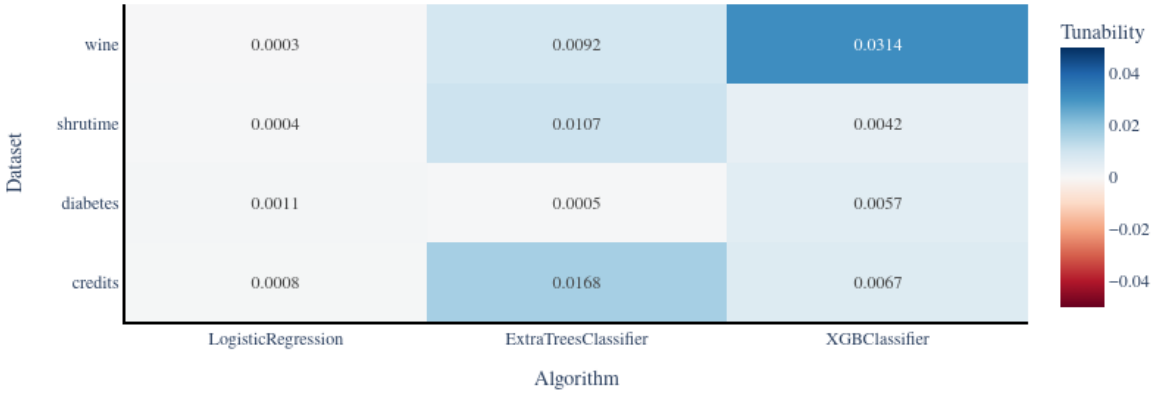
06

Tunowalność

Random Search Tunability Heatmap



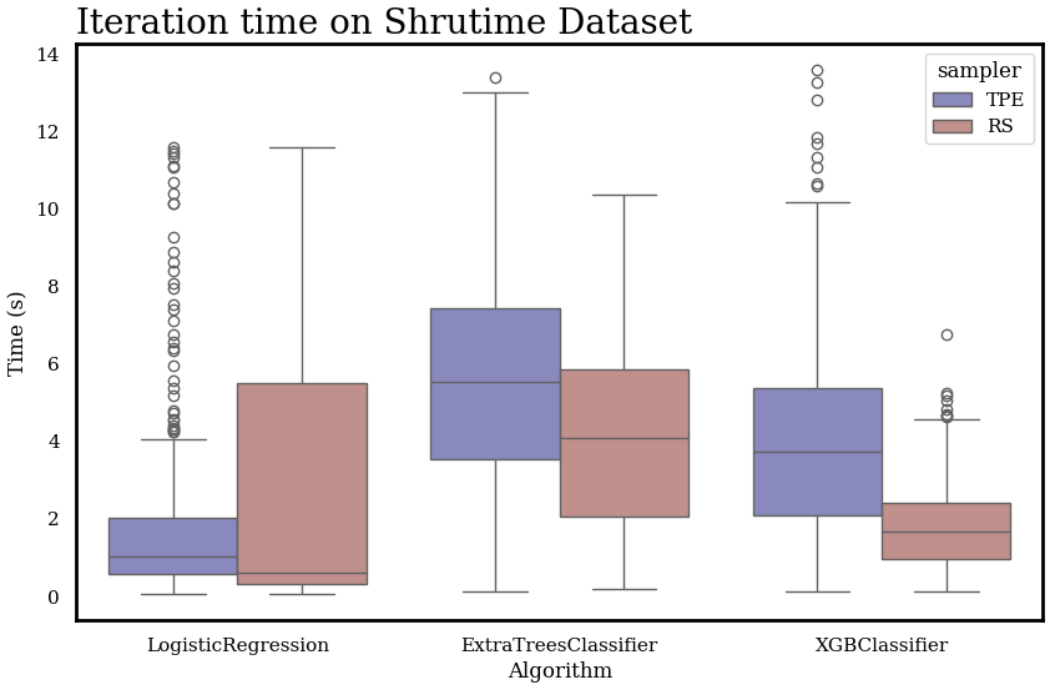
Bayesian Search Tunability Heatmap



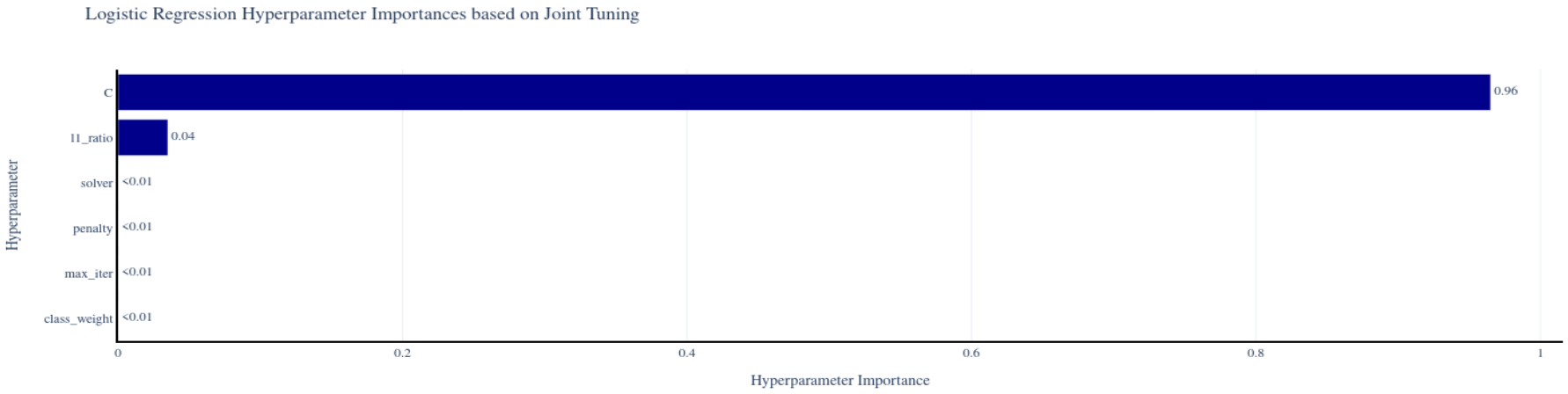
Czas strojenia - przykład dla zbioru Shrutime

Sampler	Algorytm	Czas strojenia (min)	AUC
RS	ExtraTreesClassifier	20.422524	0.8099
	LogisticRegression	15.317961	0.8328
	XGBClassifier	9.063305	0.8645
TPE	ExtraTreesClassifier	28.442875	0.8194
	LogisticRegression	9.626361	0.8328
	XGBClassifier	20.624290	0.8684

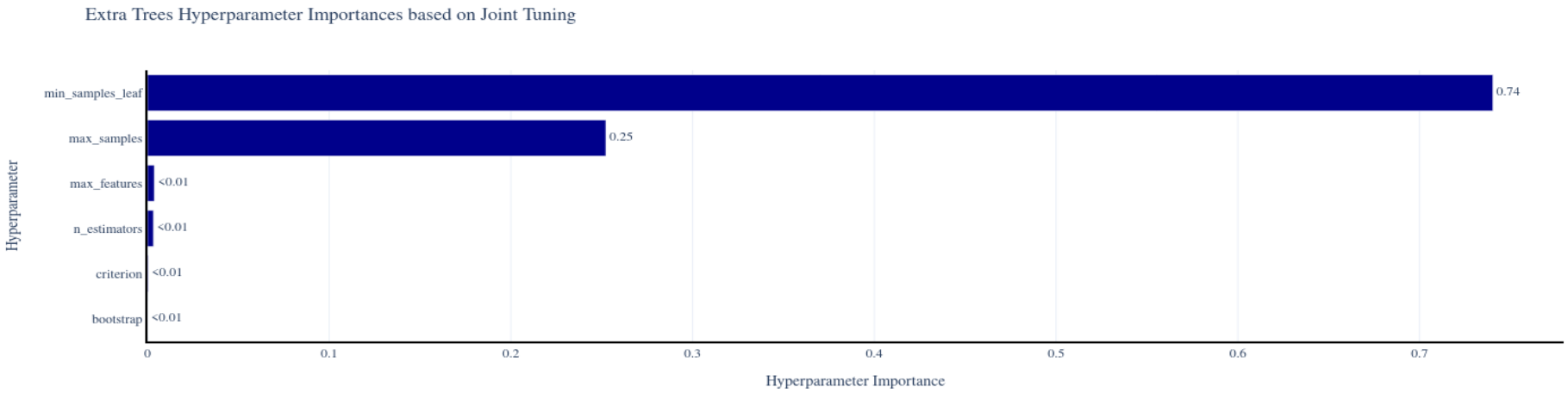
Czas pojedynczej iteracji



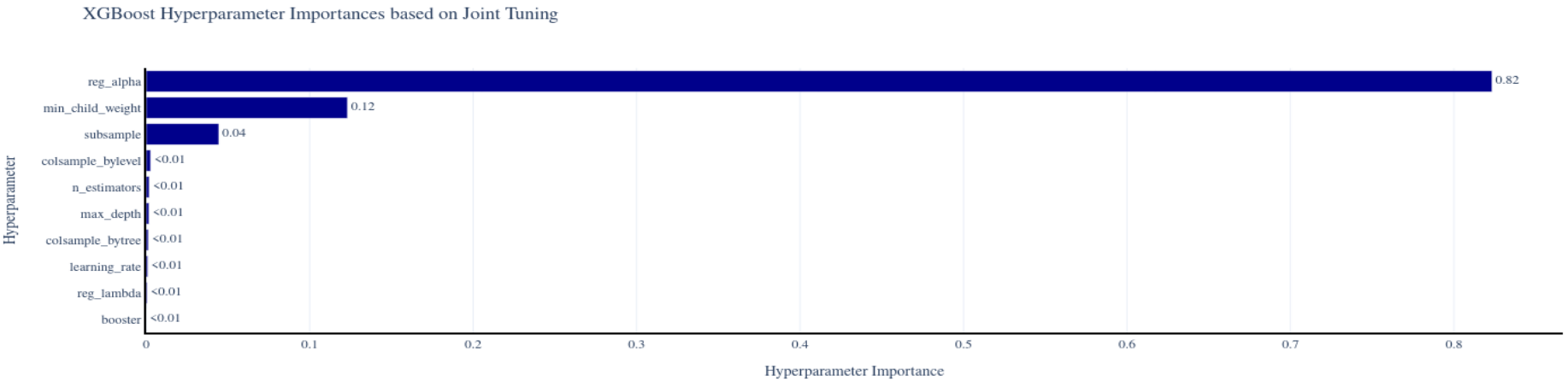
Istotność hiperparametrów



Istotność hiperparametrów



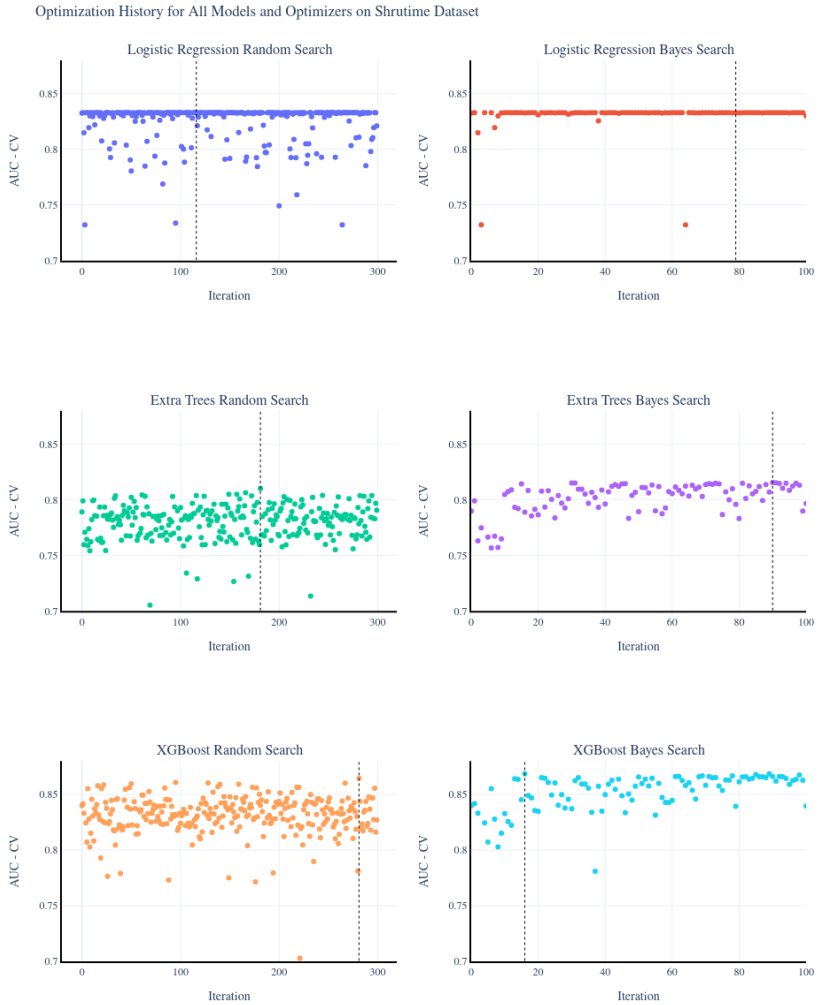
Istotność hiperparametrów



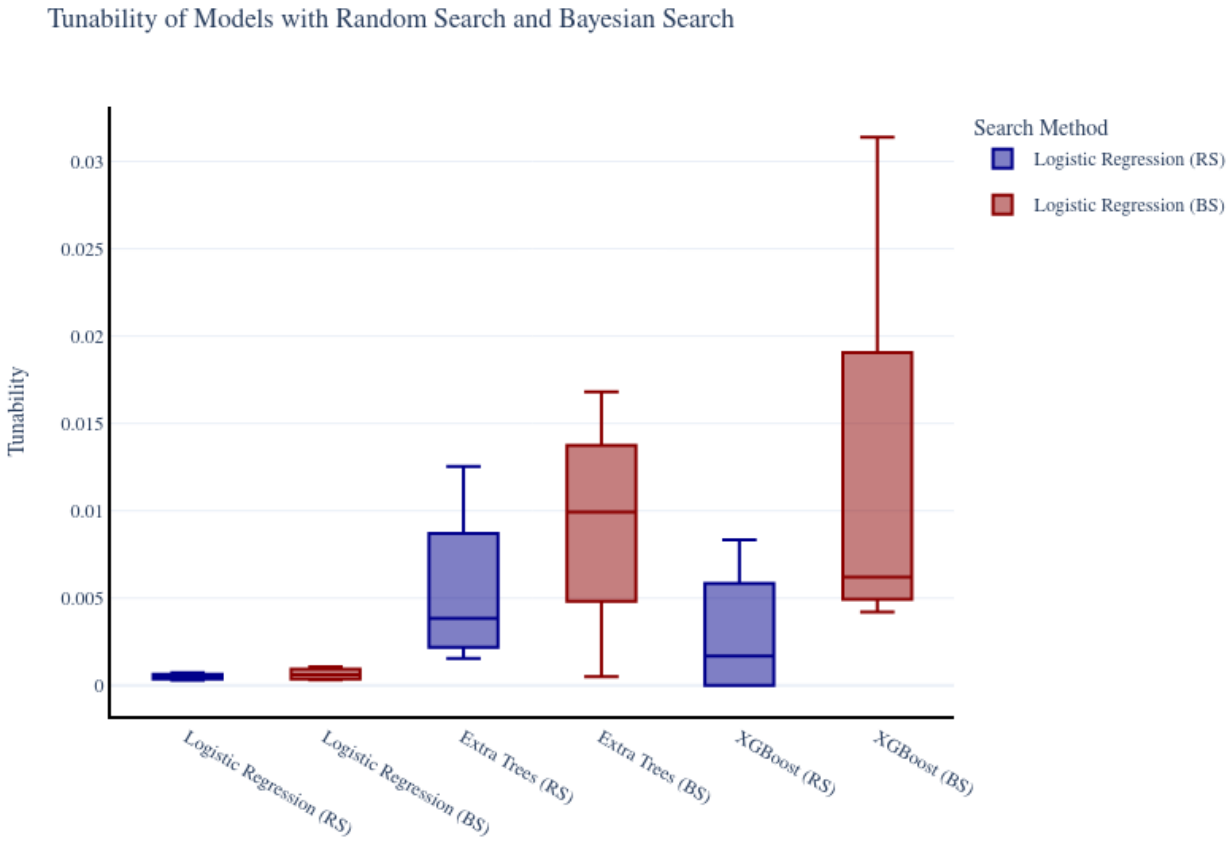
07

Bias Sampling

Historia strojenia



Wpływ techniki losowania na tunowalność



Bibliografia

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(53), 1-32.

Watanabe, S. (2023). Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv preprint arXiv:2304.11127.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

Dziękujemy za uwagę!