AutoML - dostrajalność modeli (PD1)

Damian Sarna

Wydział Matematyki i Nauk Informacyjnych

21 listopada 2023

Plan prezentacji

Dane

Wybano 4 zbiory do analizy z repozytorium OpenML

- credit-g (ID: 31, 1000 obserwacji, 61 kolumn)
- mozilla4 (ID: 1046, 15545 obserwacji, 5 kolumn)
- kc1 (ID: 1067, 2109 obserwacji, 21 kolumn)
- phoneme (ID: 1489, 5404 obserwacji, 5 kolumn)

Metody

Algorytmy ML:

- DecisionTreeClassifier (scikit-learn)
- RandomForestClassifier (scikit-learn)
- XGBoostClassifier (xgboost)

Metody optymalizacji

- RandomizedSearchCV (scikit-learn)
- BayesSearchCV (scikit-optimize)

Siatki hiperparametrów

Parametr	Od	Do	Kategorie	Uwagi
criterion	-	-	gini, entropy	-
splitter	-	-	best, random	-
max_depth	2	32	None	-
min_samples_split	-4	-1	-	log
min_impurity_decrease	-5	-1	-	log
ccp_alpha	-5	-1	-	log

Tabela: Siatka parametrów dla algorytmu DecisionTree

Siatki hiperparametrów

Parametr	Od	Do	Kategorie	Uwagi
criterion	-	-	gini, entropy, log_loss	-
n_estimators	2	500	-	-
max_depth	2	32	None	-
min_samples_split	-4	-1	-	log
max_features	-	-	sqrt, log2, None	-
min_impurity_decrease	-10	-1	-	log

Tabela: Siatka parametrów dla algorytmu RandomForest

Siatki hiperparametrów

Parametr	Od	Do	Uwagi
n_estimators	10	1000	-
max_depth	2	32	-
subsample	0.5	1	-
colsample_bytree	0.5	1	-
gamma	-10	0	log
reg_alpha	-10	2	log
reg_lambda	-10	2	log

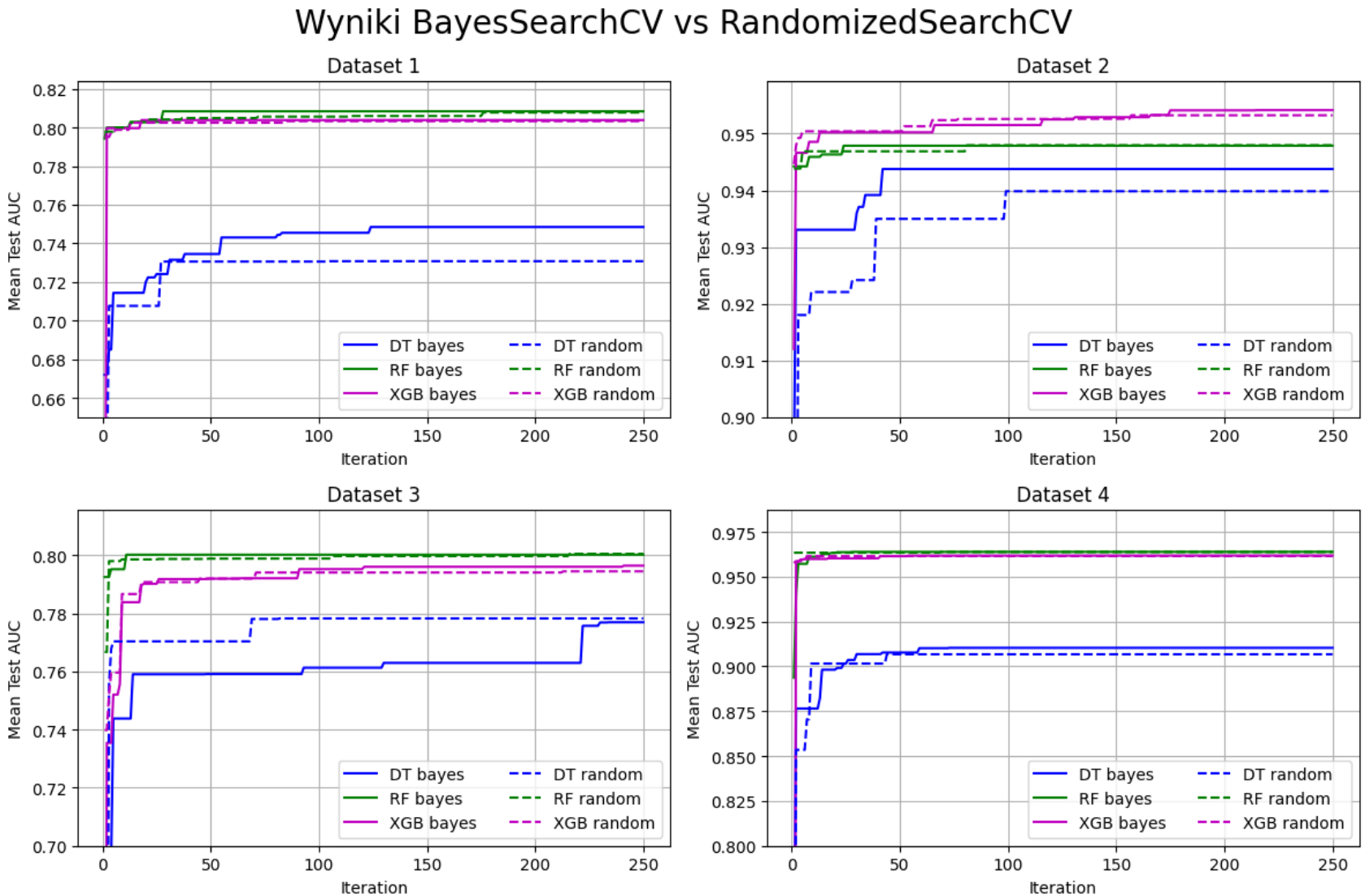
Tabela: Siatka parametrów dla algorytmu XGBoost

Procedura

Optymalizacja w 3 krokach:

- RandomizedSearchCV - 250 iteracji
- BayesSearchCV - 250 iteracji
- RandomizedSearchCV jednowymiarowo (de facto GridSearchCV) do kalibracji pojedynczych hiperparametrów

Porównanie metod kalibracji parametrów



Optymalne parametry - DT

Metoda	criterion	splitter	max_depth	min_samples_split	min_impurity_decrease	ccp_alpha	AUC
Random	entropy	best	13	0.0464	0.0006	0.0046	0.8202
Bayes	gini	best	9	0.1	0.0001	0.0037	0.8363

Tabela: Optymalne hiperparametry algorytmu DecisionTree

Optymalne parametry - RF

Metoda	criterion	n_est	max_depth	min_samples	max_features	min_imp_de	AUC
Random	log_loss	232	15	0.001	log2	0.0	0.8708
Bayes	entropy	500	24	0.0004	sqrt	0.0	0.8783

Tabela: Optymalne parametry algorytmu RandomForest

Optymalne parametry - XGB

Metoda	n_est	max_depth	subsample	colsample_bytree	gamma	reg_alpha	reg_lambda	AUC
Random	520	13	0.94	0.67	1.0	1.0	0	0.8683
Bayes	1000	17	0.83	0.5	0	5.42	0.0	0.8752

Tabela: Optymalne parametry algorytmu XGBoost

Dostrajalność modelu

Dostrajalność modelu na j-tym zbiorze definiujemy jako:

$$d^j = AUC(\theta^{(j)*}) - AUC(\theta^*), \tag{1}$$

gdzie $\theta^{(j)*}$ to optymalny wektor hiperparametrów dla zbioru j, a θ^* to globalny optymalny wektor parametrów.

Algorytm	Dane 1	Dane 2	Dane 3	Dane 4	Średnia dostrajalność
DecisionTree	0.0055	0.0122	0.0284	0.0285	0.0187
RandomForest	0.0031	0.0015	0.0291	0.0032	0.0092
XGBoost	0.0036	0.0058	0.0194	0.0106	0.0098

Tabela: Dostrajalność testowanych algorytmów

Dostrajalność hiperparametrów

Dostrajalność i-tego parametru na j-tym zbiorze definiujemy jako:

$$d_i^{(j)} = AUC(\theta_i^{(j)*}) - AUC(\theta^*), \tag{2}$$

gdzie $\theta_i^{(j)*}$ to optymalny wektor hiperparametrów dla zbioru j, który różni się od θ^* co najwyżej na i-tym miejscu.

Dostrajalność hiperparametrów

criterion	splitter	max_depth	min_samples_split	min_impurity_decrease	ccp_alpha
0.0010	0.0030	0.0027	0.0020	0.0014	0.0037

Tabela: Dostrajalność hiperparametrów algorytmu DecisionTree

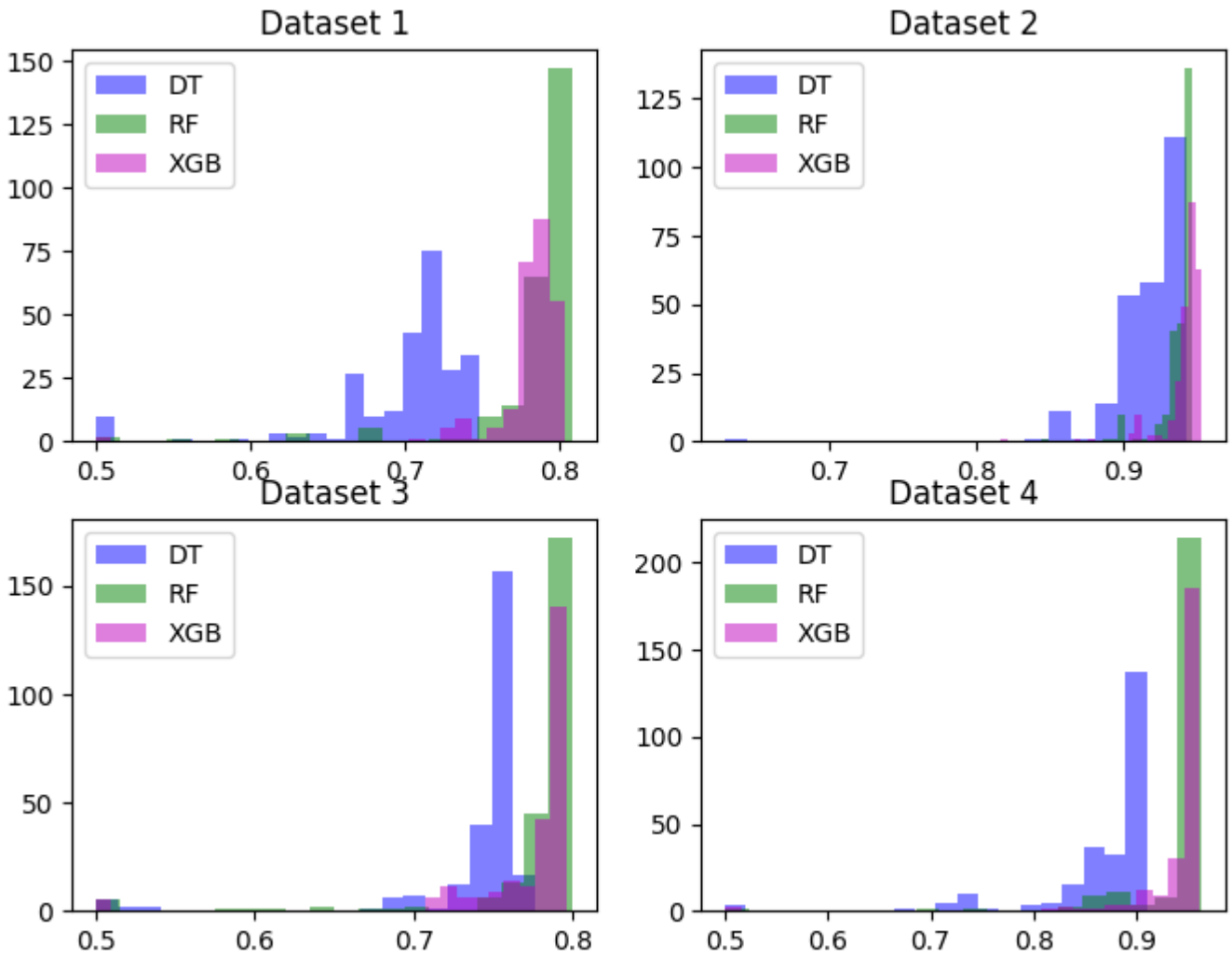
criterion	n_estimators	max_depth	min_samples_split	max_features	min_impurity_decrease
0	0.0007	0.0075	0.0063	0.0	0.0076

Tabela: Dostrajalność hiperparametrów algorytmu RandomForest

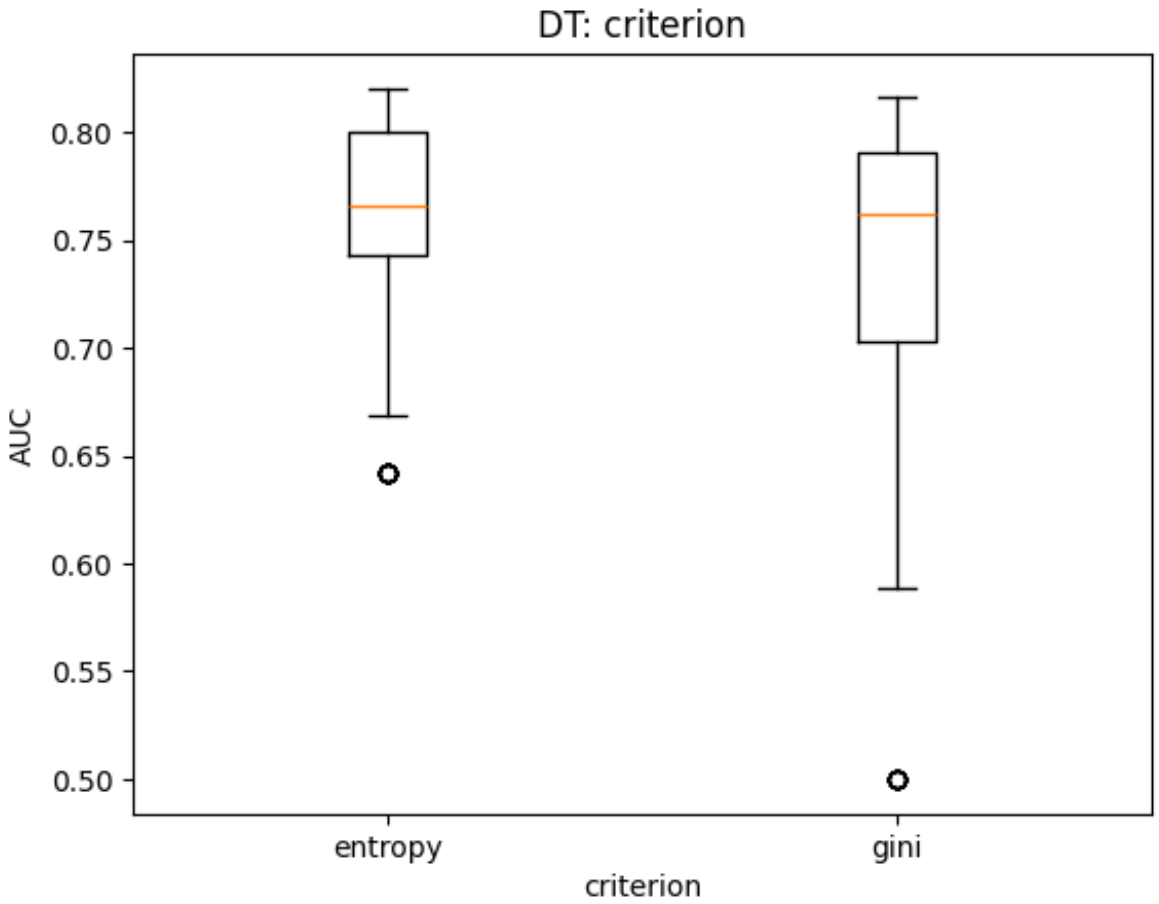
n_estimators	max_depth	subsample	colsample_bytree	gamma	reg_alpha	reg_lambda
0.0006	0.0037	0.0013	0.0030	0.0019	0.0073	0.0059

Tabela: Dostrajalność hiperparametrów algorytmu XGBoost

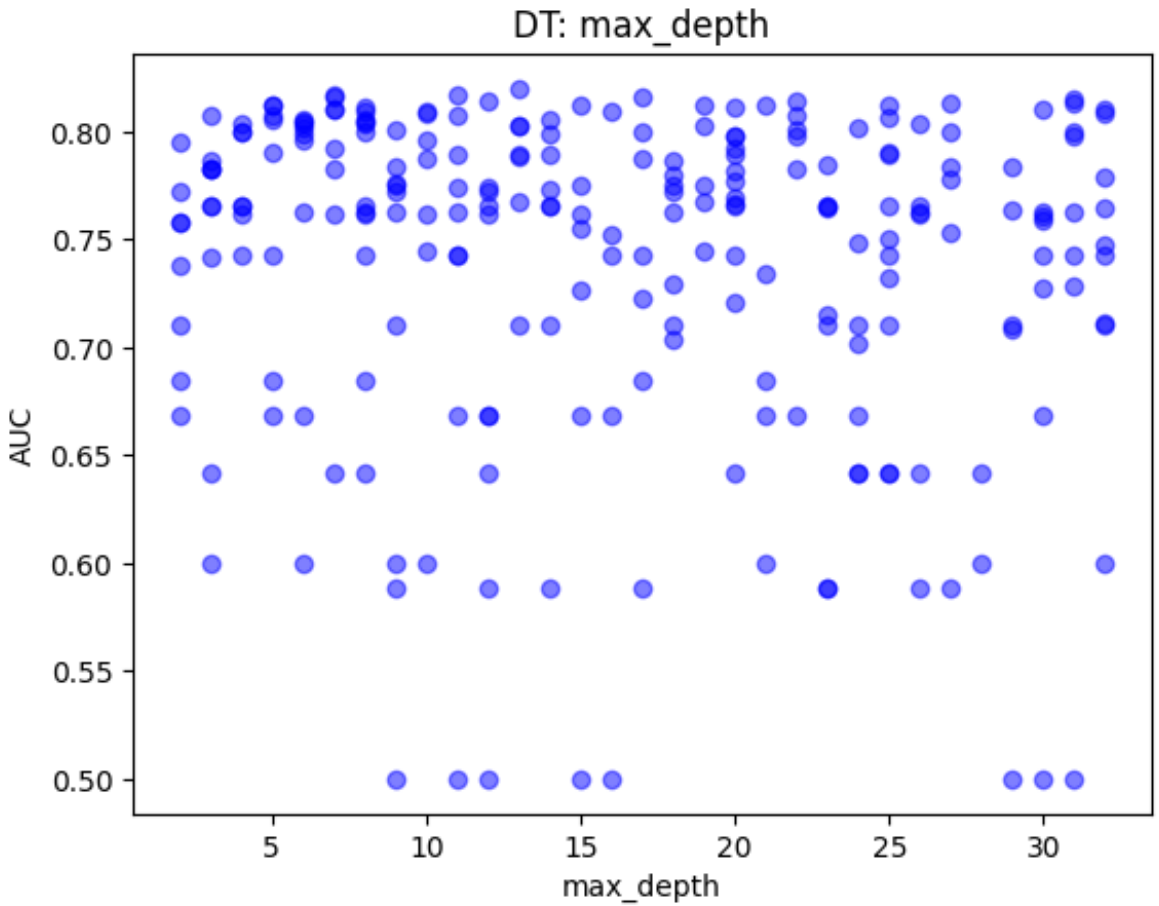
Porównanie AUC modeli



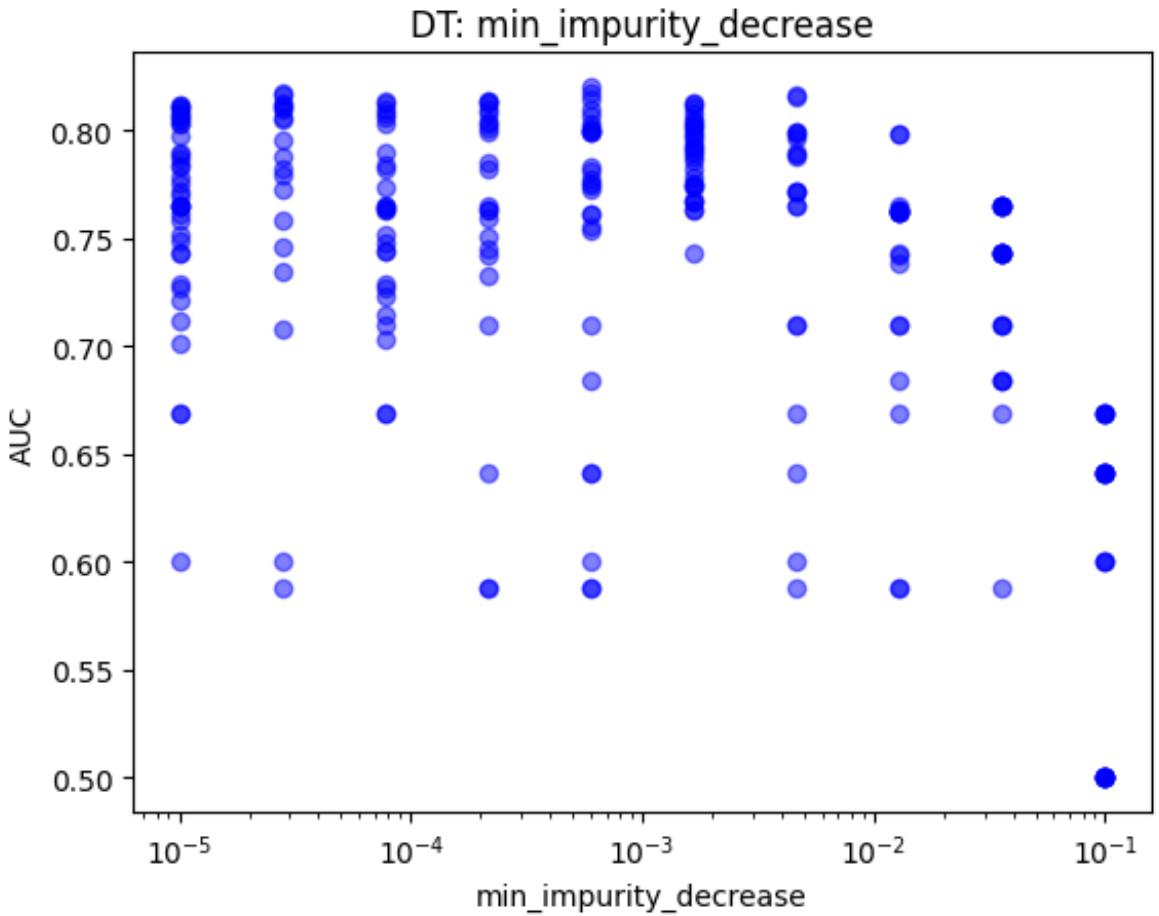
DT - parametry domyślne



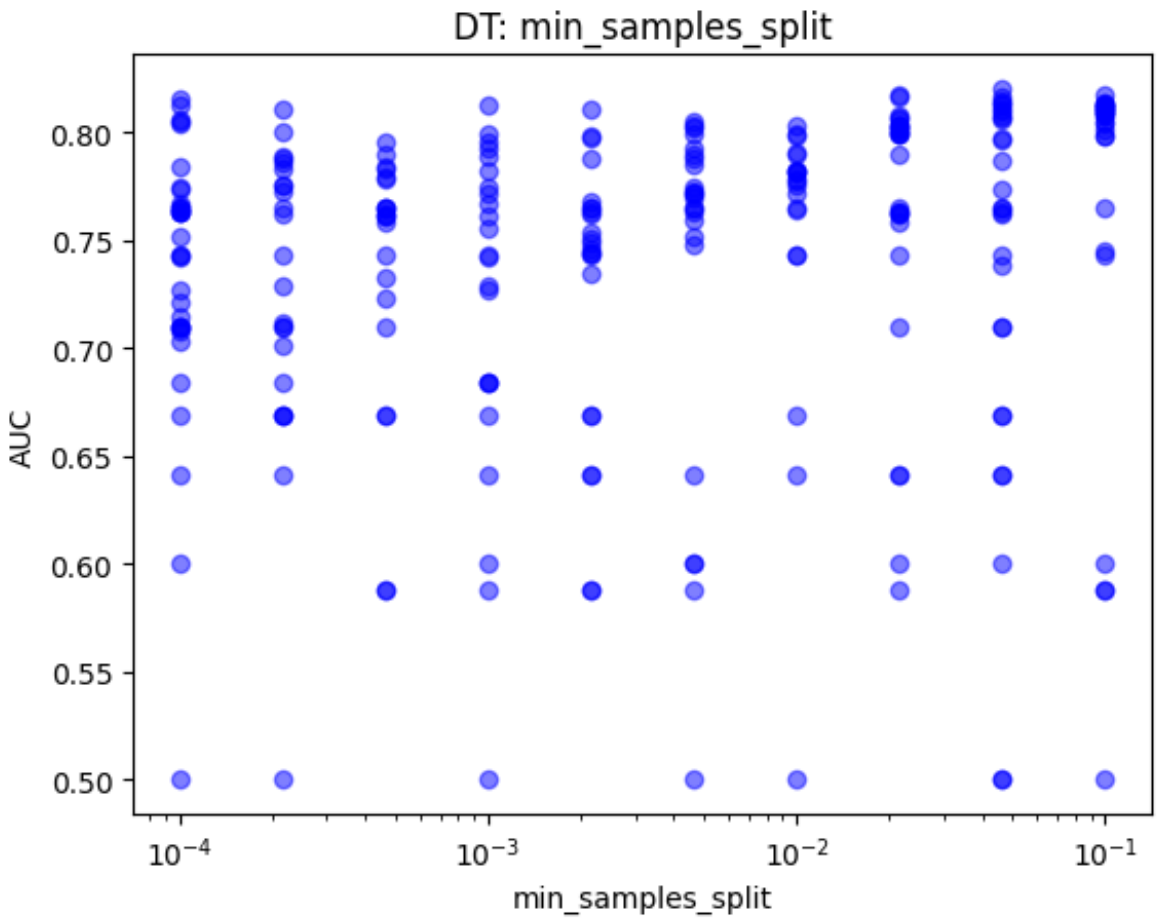
DT - parametry domyślne



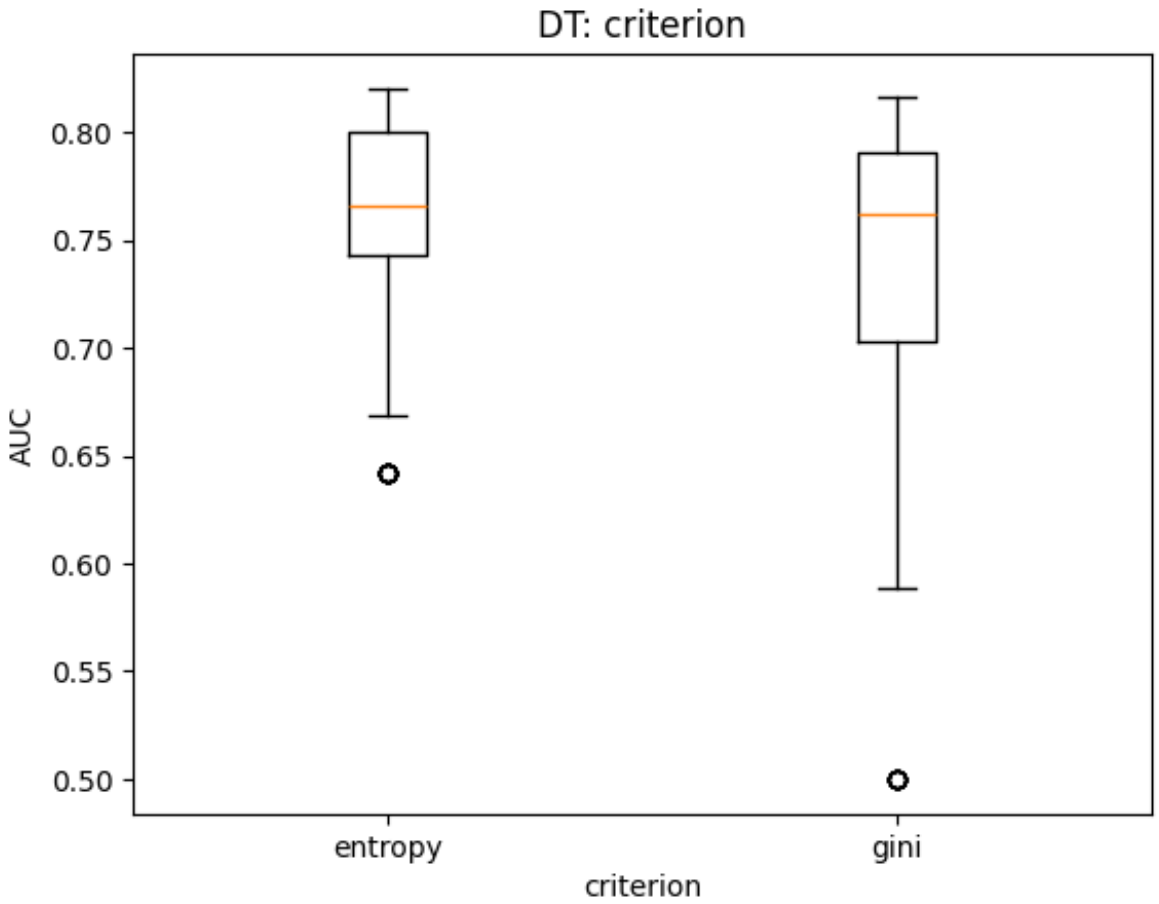
DT - parametry domyślne



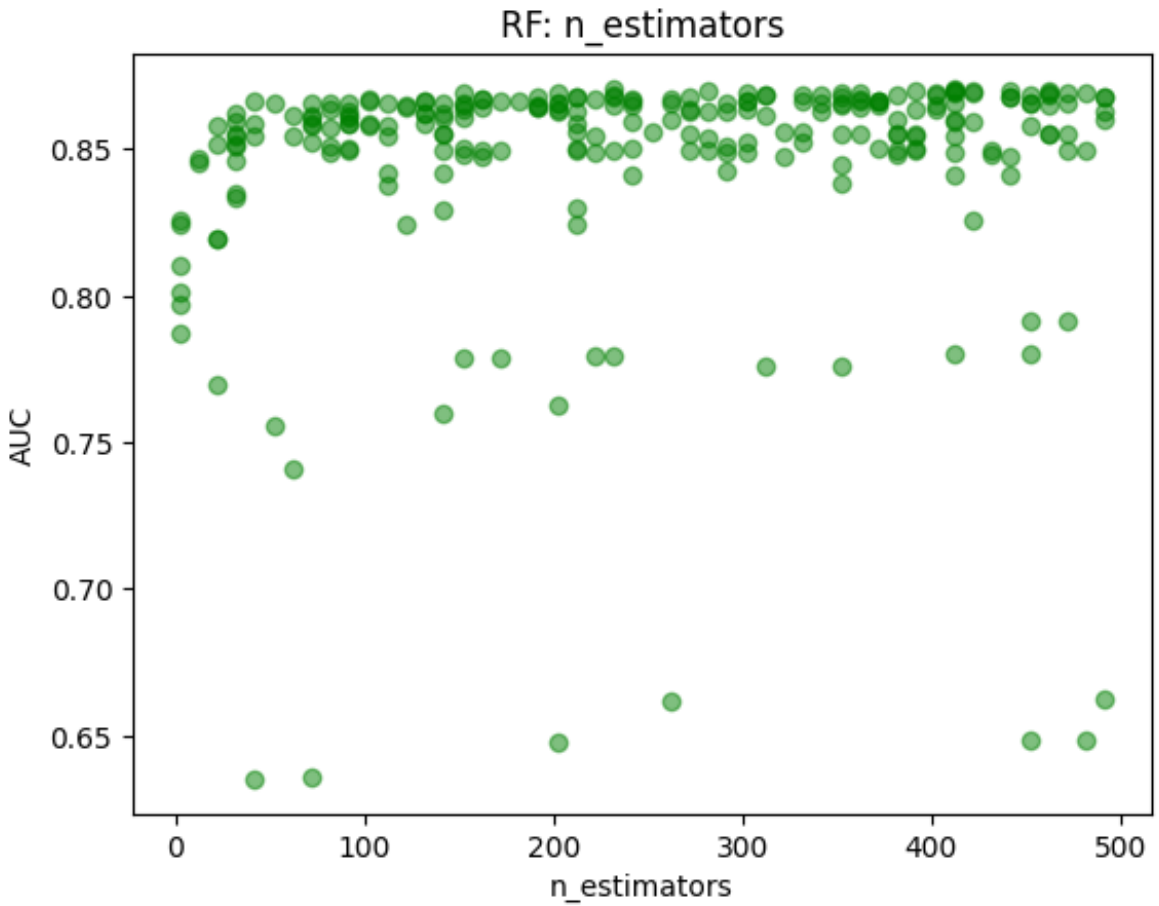
DT - parametry domyślne



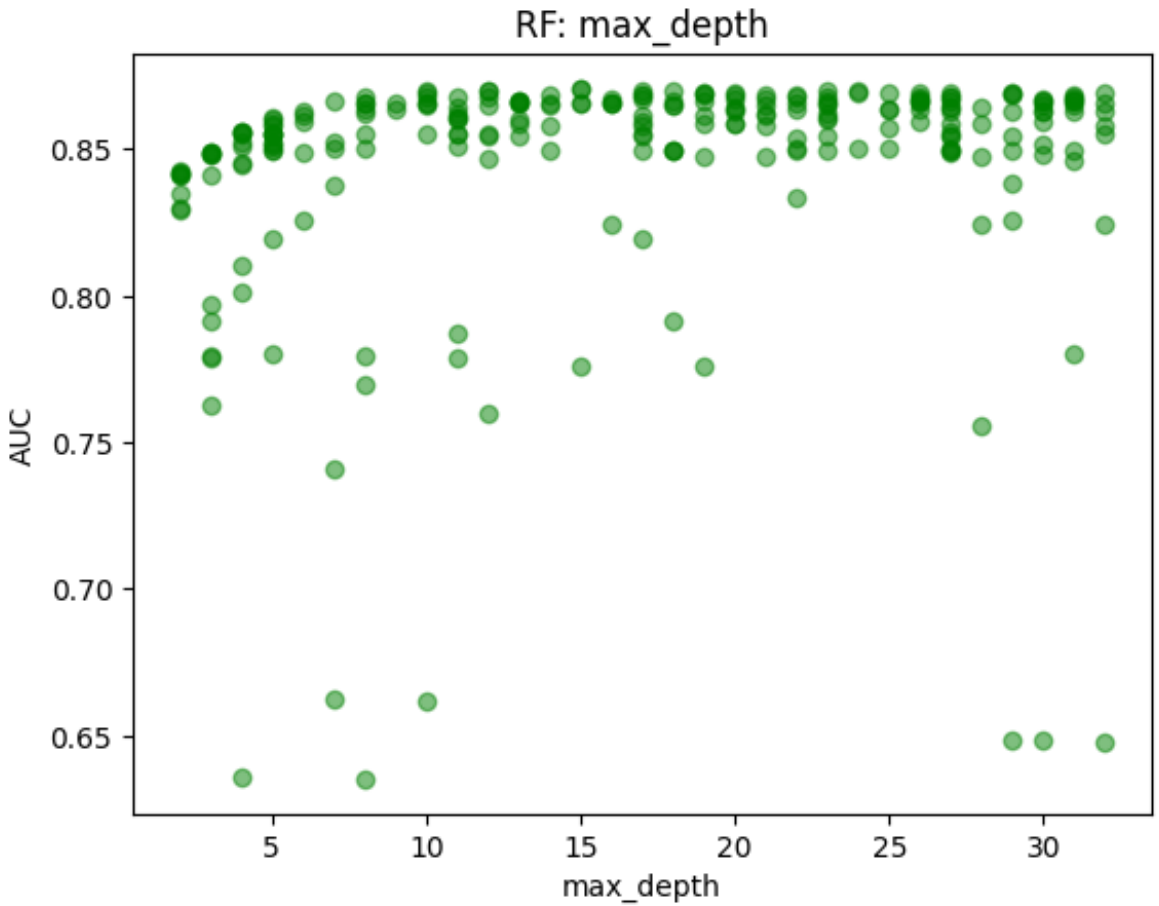
DT - parametry domyślne



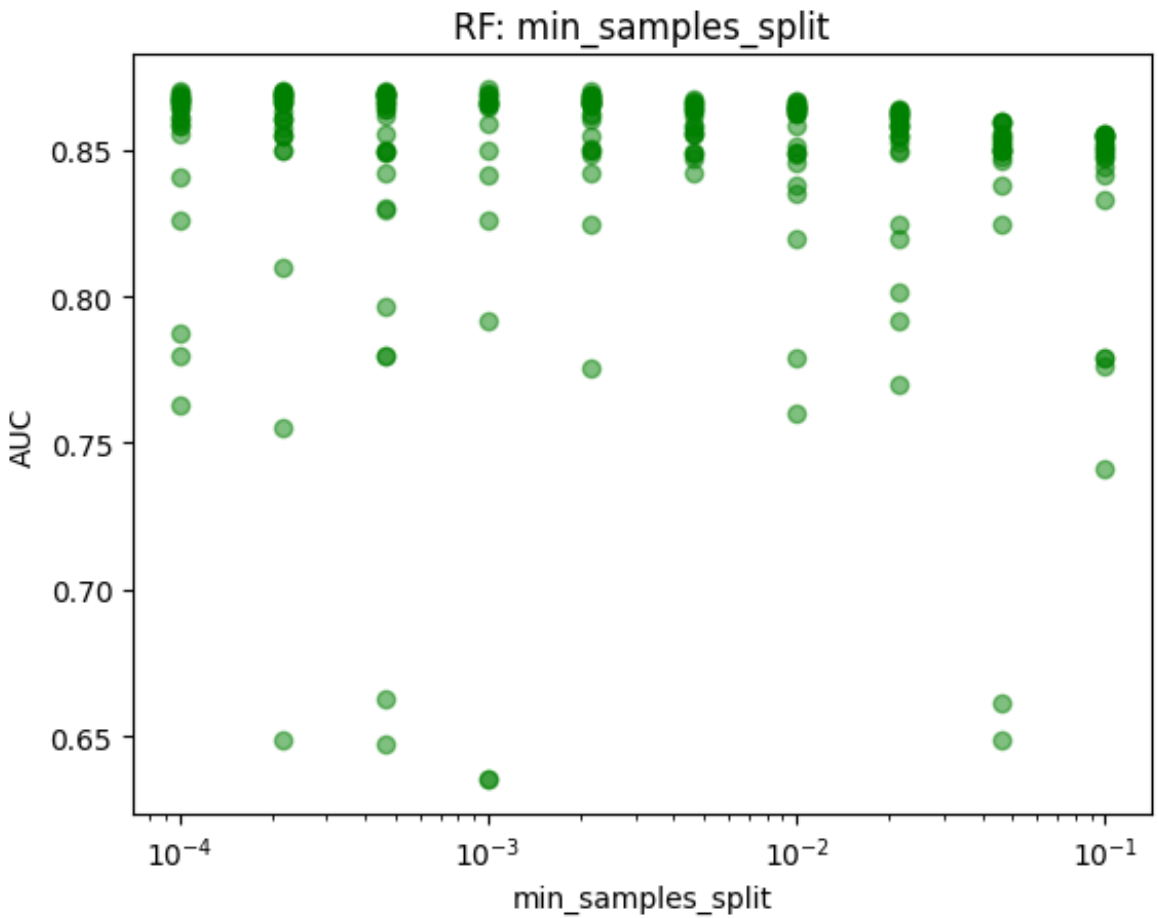
RF - parametry domyślne



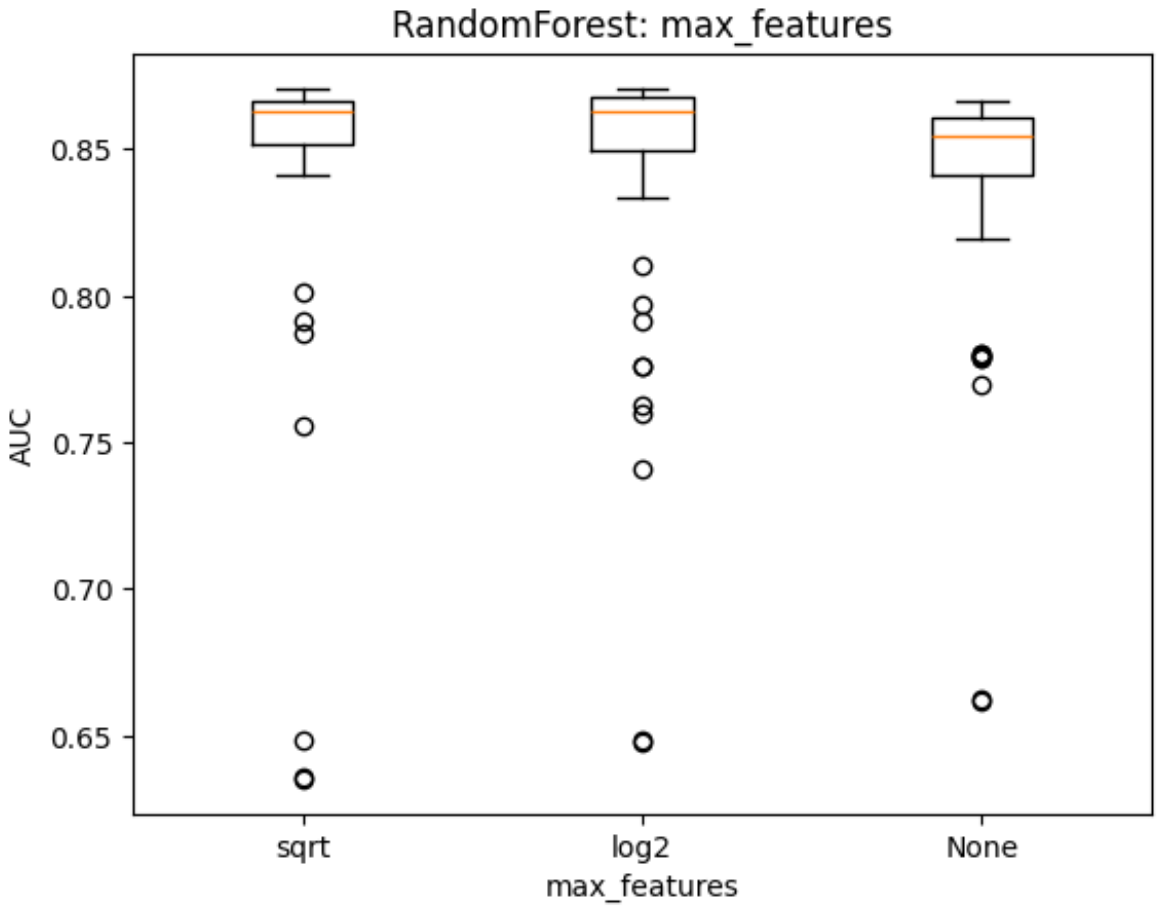
RF - parametry domyślne



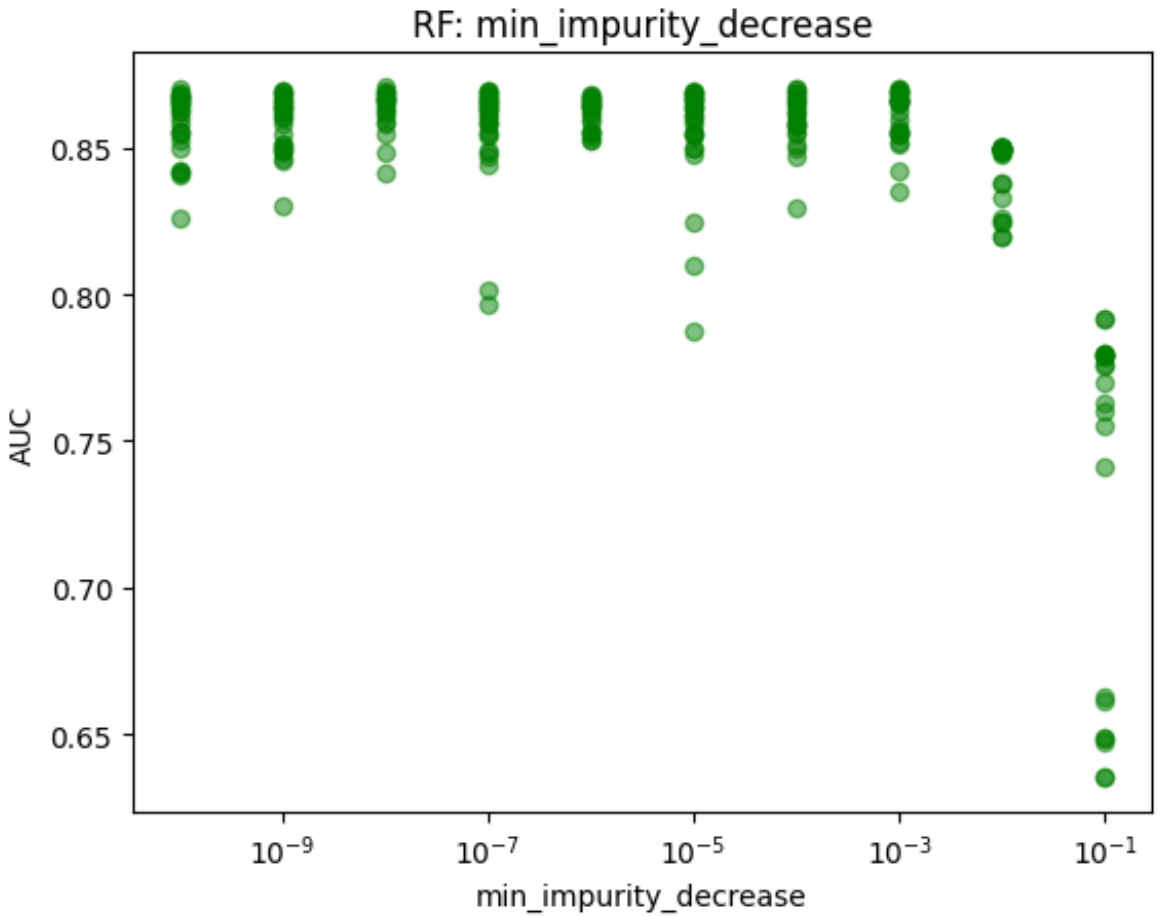
RF - parametry domyślne



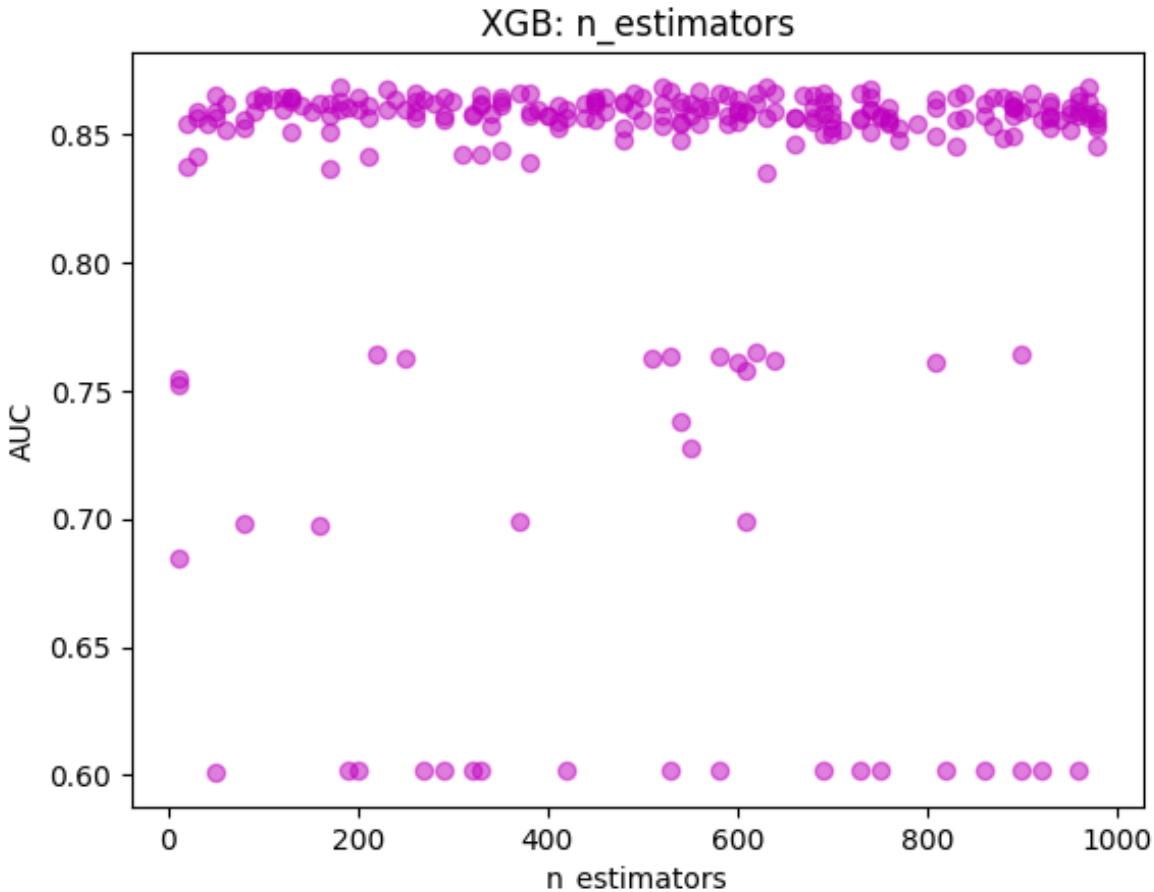
RF - parametry domyślne



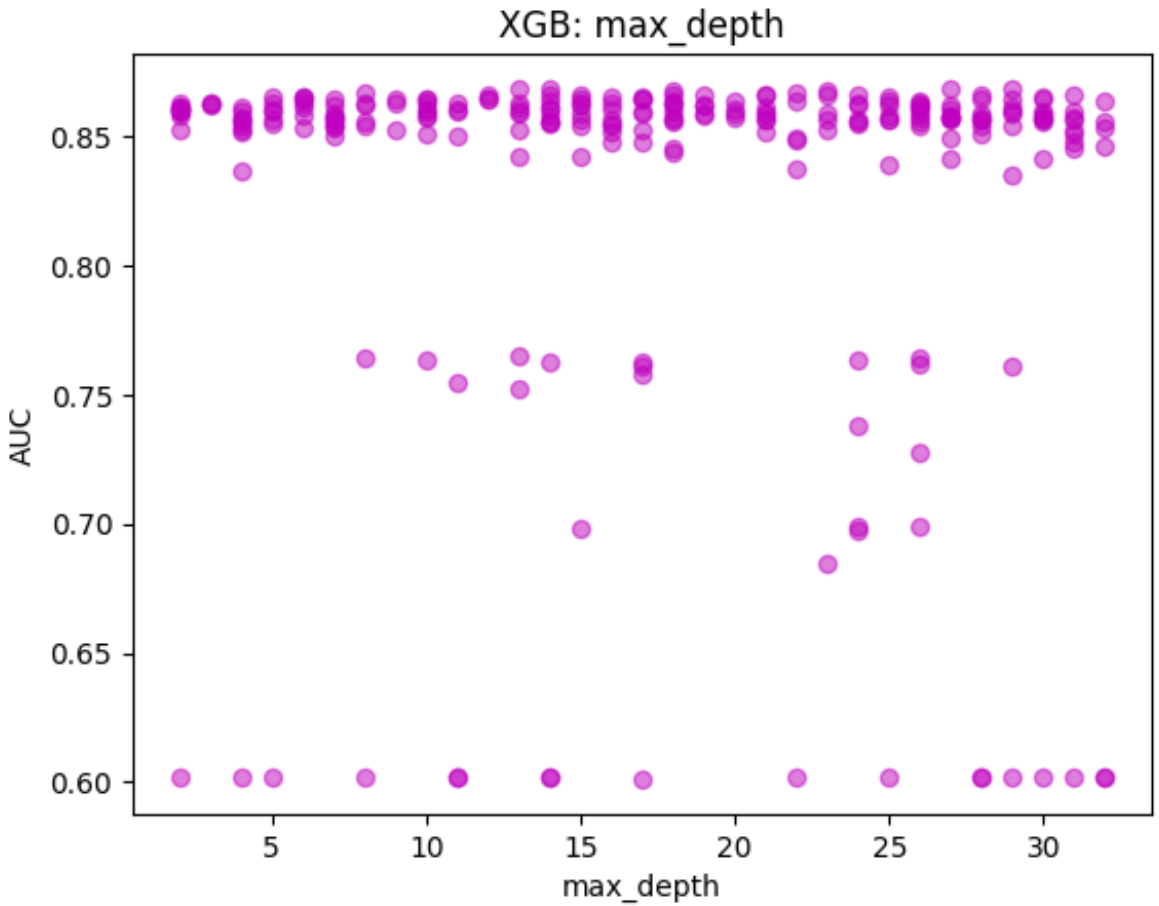
RF - parametry domyślne



XGB - parametry domyślne



RF - parametry domyślne



RF - parametry domyślne