Tunowalność wybranych algorytmów uczenia maszynowego

Katsiaryna Bokhan, Monika Jarosińska

MiNI PW

06.11.2024

Plan prezentacji

1. Wstęp do eksperymentu

2. Wizualizacja otrzymanych wyników
   - Tunowalność SVM, CatBoost, ExtraTrees względem różnych metryk
   - Porównanie z algorytmami o domyślnych hiperparametrach
   - Analiza metod samplingu pod względem liczby iteracji

3. Test Manna-Whitneya

Zbiory danych - krótki opis

1. Zbiór danych 1 (833) - syntetyczny, celem jest przewidywanie, czy **klient w banku został obsłużony** w zależności od stanu kolejki i lokalizacji oddziału banku
2. Zbiór danych 2 (44157) - informacje o **ruchach oczu** (liczba fiksacji, czas trwania fiksacji, pozycja wzroku itp) uczestników czytających zdania, aby określić, na ile słowa są istotne dla odpowiedzi na zadane pytanie.
3. Zbiór danych 3 (1120) - dane służące do klasyfikacji wysokoenergetycznych **cząstek gamma i cząstek tła** pochodzących z promieni kosmicznych na podstawie obrazów promieniowania Czerenkowa.
4. Zbiór danych 4 (45553) - zbiór danych dotyczy **oceny zdolności kredytowej** pożyczkobiorców na podstawie ich profilu finansowego.

Zbiory danych - preprocessing

Zbiór danych (ID)	Liczba rekordów	Rozkład zmiennej objaśnianej	Liczba kolumn
Zbiór danych 1 (833)	8192	31% vs 69%	32
Zbiór danych 2 (44157)	7608	50% vs 50%	23
Zbiór danych 3 (1120)	19020	35% vs 65%	10
Zbiór danych 4 (45553)	9871	48% vs 52%	23

Tabela: Informacje na temat zbiorów danych.

- Żaden zbiór nie zawierał braków danych
- Postanowiono, że wprowadzimy tylko niezbędne zmiany w danych
- Jeden pipeline:
- **OneHotEncoder** - transformacja zmiennych kategorycznych
- **MinMaxScaler** - transformacja zmiennych numerycznych

Wybrane algorytmy i metody samplingu do analizy

Wybrane algorytmy:

- **Support Vector Machine**
- **ExtraTreesClassifier**
- **CatBoostClassfifier**

Wybrane metody samplingu:

- **RandomSearch** - własna klasa *RandomSearchWithMetrics*
- **Bayes Optimization** - funkcja *skopt.BayesSearchCV* z biblioteki *scikit-optimize*

Wybrane metryki do optymalizacji:

- ***ROC AUC*** - główna metryka
- *BrierScore*, *Accuracy* - dodatkowo obliczone metryki dla RandomSearch

Ustalone siatki hiperparametrów dla **SVM i CatBoost**

Hiperparametr	Możliwe wartości	Liczba wartości
`kernel`	'linear', 'poly', 'rbf', 'sigmoid'	4
`C`	$2^{-10}, 2^{-9}, \ldots, 2^{9}$	20
`gamma`	$2^{-10}, 2^{-9}, \ldots, 2^{9}$	20
`degree`	2, 3, 4, 5	4
Liczba możliwych kombinacji		6400

Tabela: Siatka hiperparametrów dla **SVM**.

Hiperparametr	Możliwe wartości	Liczba wartości
`iterations`	2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 15, 18, 21, 25, 29, 34, 40, 47, 56, 65, 77, 90, 105, 124, 145, 170, 200	27 (log space used)
`learning_rate`	0.01, 0.1325, 0.255, 0.3775, 0.5	5
`depth`	1, 3, 5, 7, 9, 11, 13, 16	8
`l2_leaf_reg`	1.0, 3.0, 5.0, 7.0, 9.0, 11.0, 13.0, 15.0	8
`colsample_bylevel`	0.1, 0.325, 0.55, 0.775, 1.0	5
Liczba możliwych kombinacji		43200

Tabela: Siatka hiperparametrów dla **CatBoostClassifier**.

Ustalona siatka hiperparametrów dla **ExtraTreesClassifier**

Hiperparametr	**Możliwe wartości**	**Liczba wartości**
`n_estimators`	2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 19, 21, 25, 29, 33, 38, 44, 51, 58, 67, 78, 89, 103, 119, 137, 158, 182, 209, 241, 277, 319, 368, 424, 488, 562, 647, 745, 858, 988, 1137, 1310, 1508, 1737, 2000	46 (log space used)
`max_depth`	None, 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 30	16
`min_samples_split`	2, 3, 4, 5, 6, ... 58, 59, 60	59
`min_samples_leaf`	1, 2, 3, 4, 5, ..., 58, 59, 60	60
`min_weight_fraction_leaf`	20 values from 0 to 0.5 uniformly distributed	20
`max_leaf_nodes`	None, 2, 3, 4, ... , 56, 57, 58, 60	59
`max_features`	None, "sqrt", "log2", 20 values from 0.1 to 1 uniformly distributed	23
`criterion`	'gini', 'entropy'	2
Liczba możliwych kombinacji		141,423,283,200

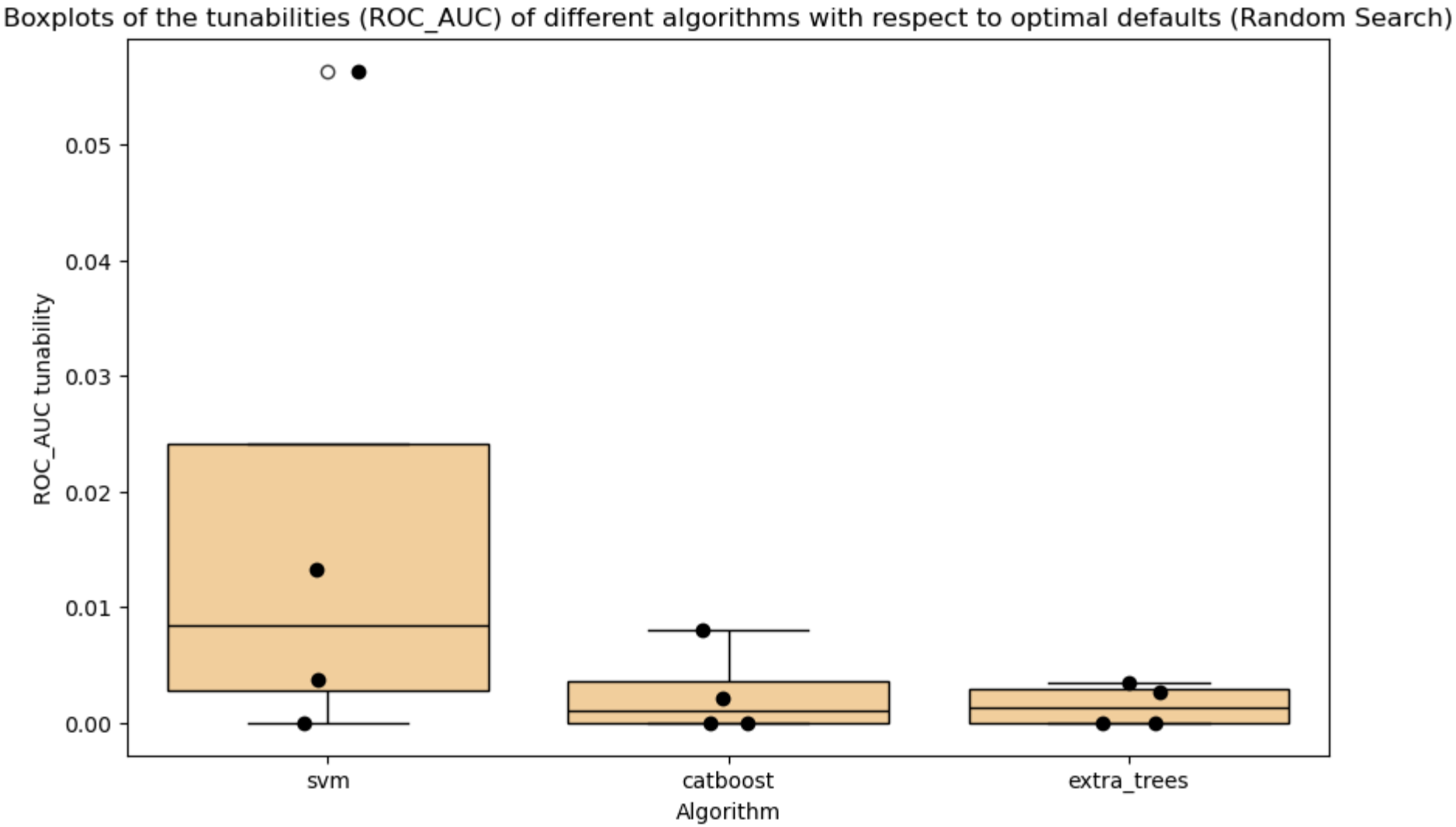
Tabela: Siatka hiperparametrów dla **ExtraTreesClassifier**.

Liczba iteracji każdej metody samplingu dla każdego algorytmu

Zbiór danych	**RandomSearch**			**BayesOptimization**		
	SVM	**CatBoost**	**ExtraTrees**	**SVM**	**CatBoost**	**ExtraTrees**
Zbiór danych 1	449 (2)	449 (1)	449 (5)	100 (1)	100 (1)	100 (1)
Zbiór danych 2	449 (2)	449 (1)	449 (5)	200 (1)	200 (1)	200 (1)
Zbiór danych 3	449 (2)	449 (1)	449 (5)	100 (1)	100 (1)	100 (1)
Zbiór danych 4	449 (2)	449 (1)	449 (5)	100 (1)	100 (1)	100 (1)

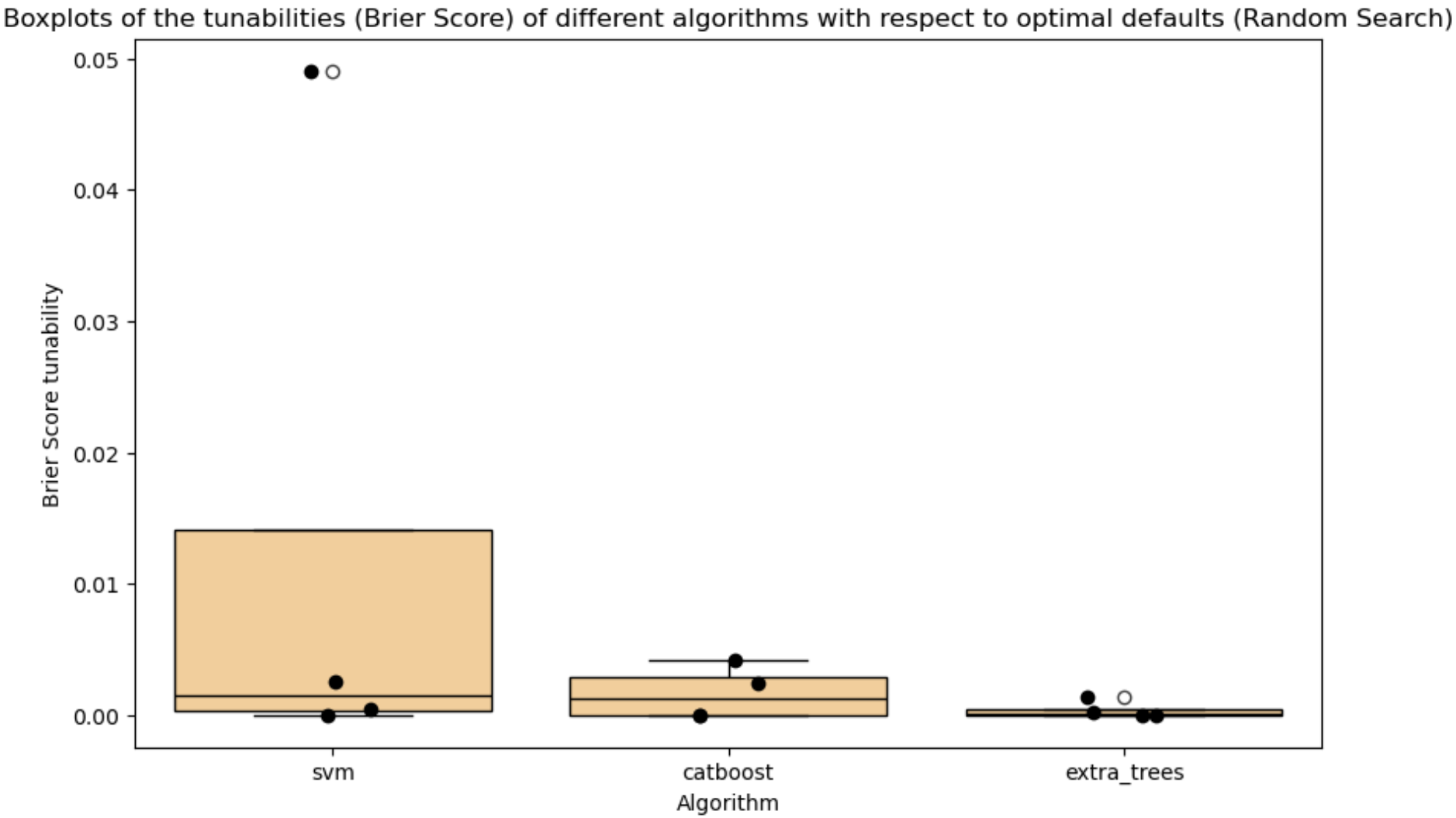
Tabela: Liczba iteracji dla metod **BayesOptimization** oraz **RandomSearch**. W nawiasach znajduję się liczba powtórzeń pięciokrotnej kroswalidacji.

Tunowalność **SVM, CatBoost, ExtraTrees** względem ROC AUC



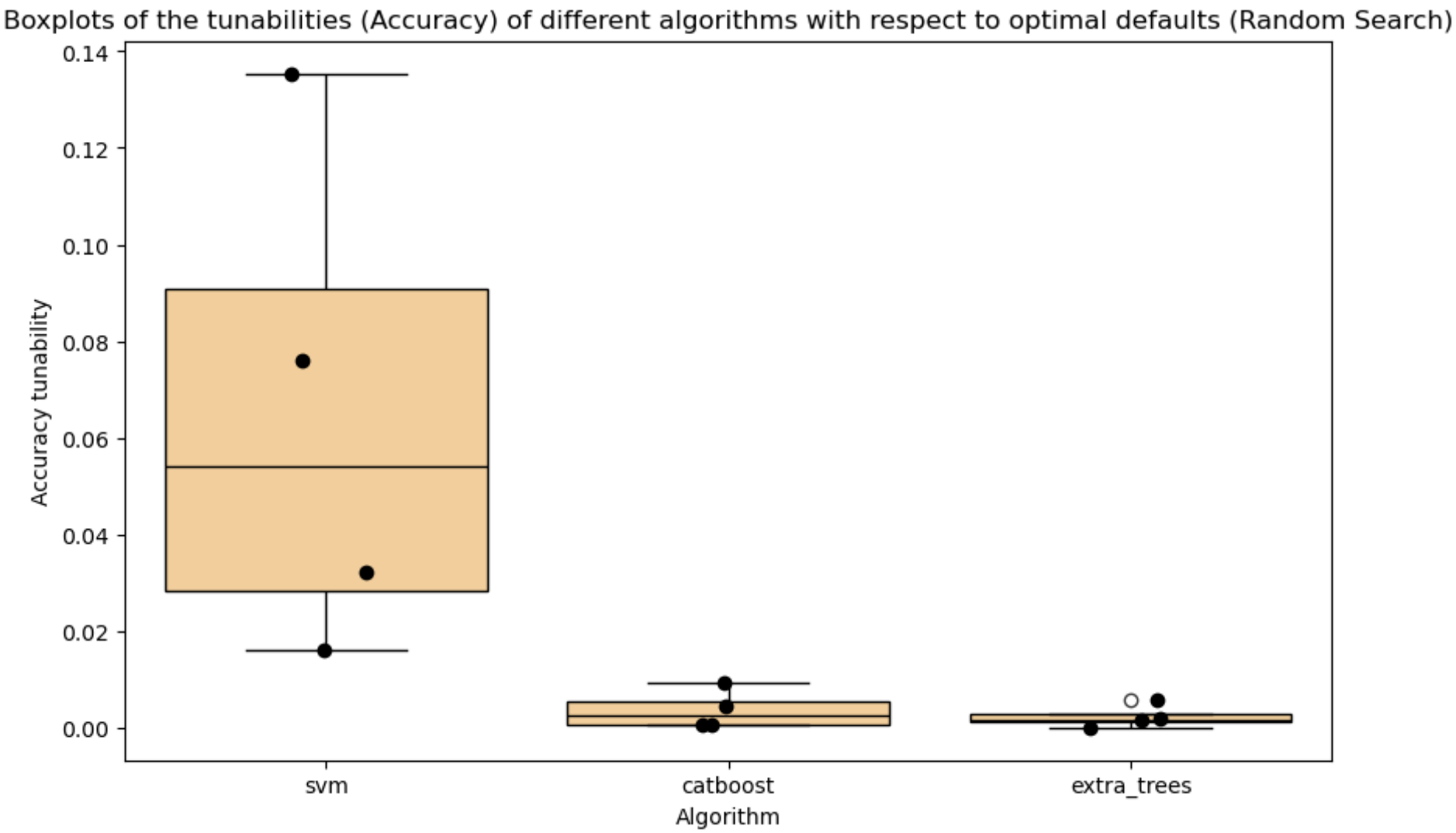
Rysunek: Tunowalność modelu **SVM** przy użyciu metryki *ROC AUC*.

Tunowalność **SVM, CatBoost, ExtraTrees** względem 1−Brier Score



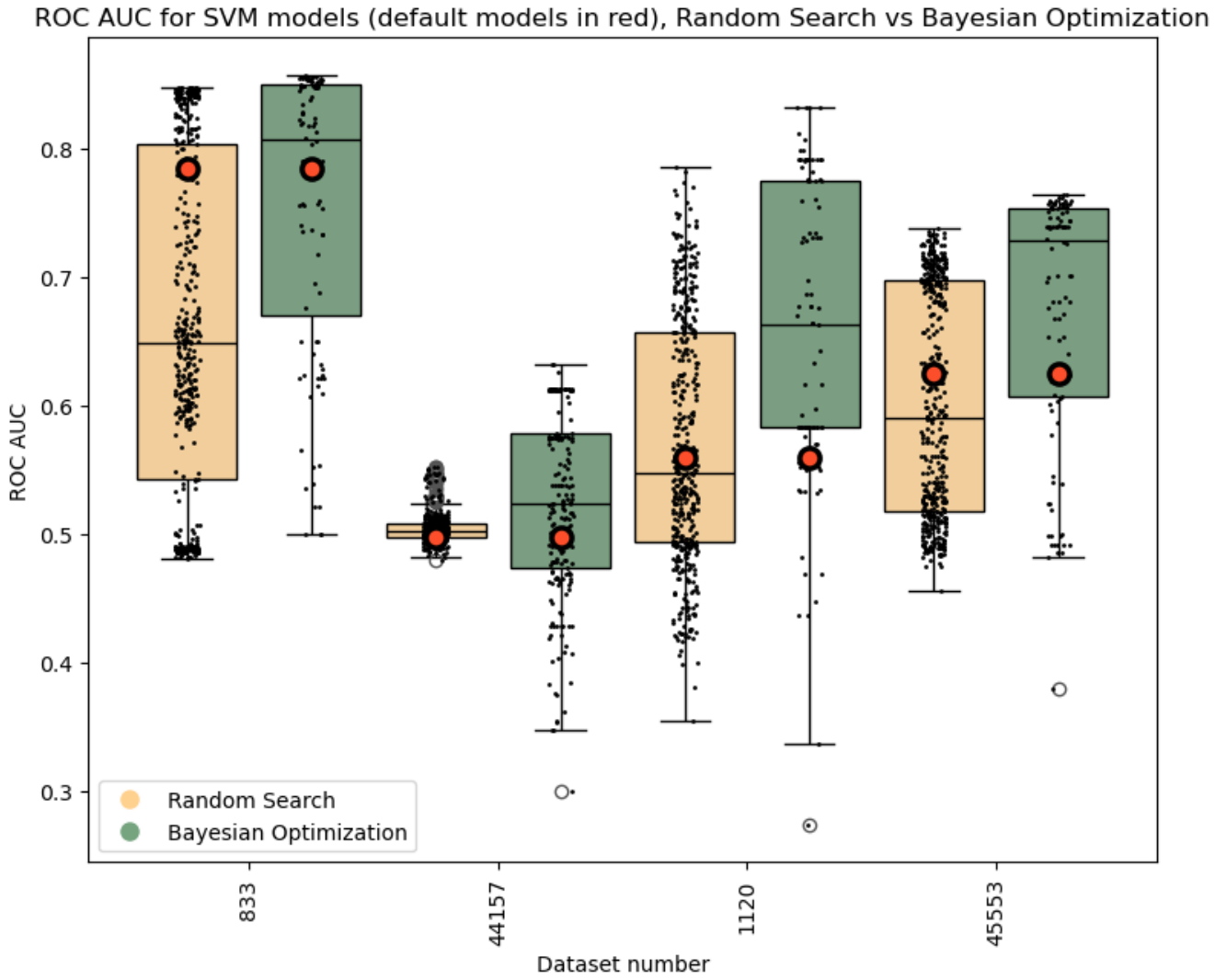
Rysunek: Tunowalność modelu **CatBoostClassifier** przy użyciu metryki 1−*Brier Score*.

Tunowalność **SVM, CatBoost, ExtraTrees** względem Accuracy



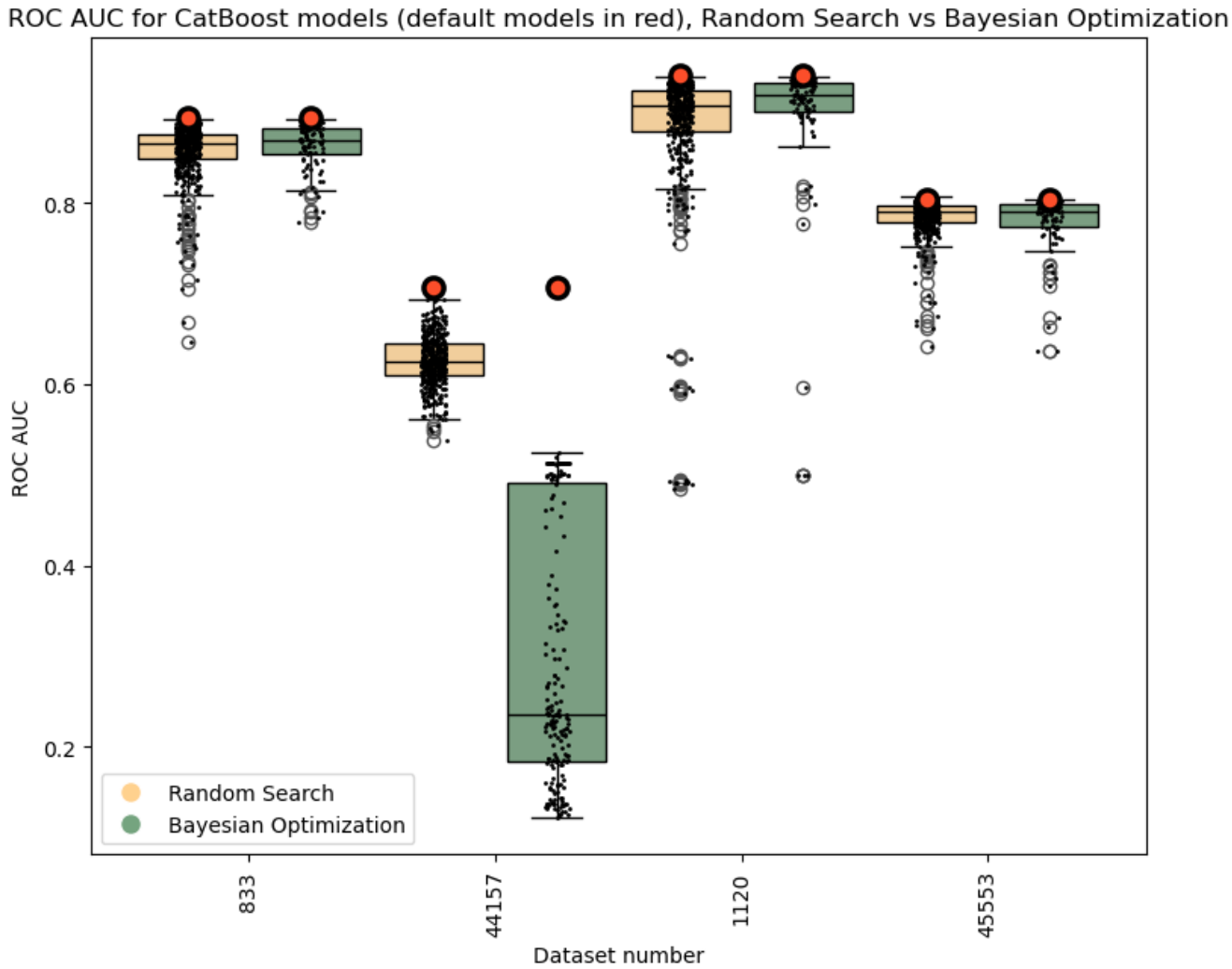
Rysunek: Tunowalność modelu **ExtraTreesClassifier** przy użyciu metryki *Accuracy*.

Metody **BayesOptimization** oraz **RandomSearch** dla **SVM**



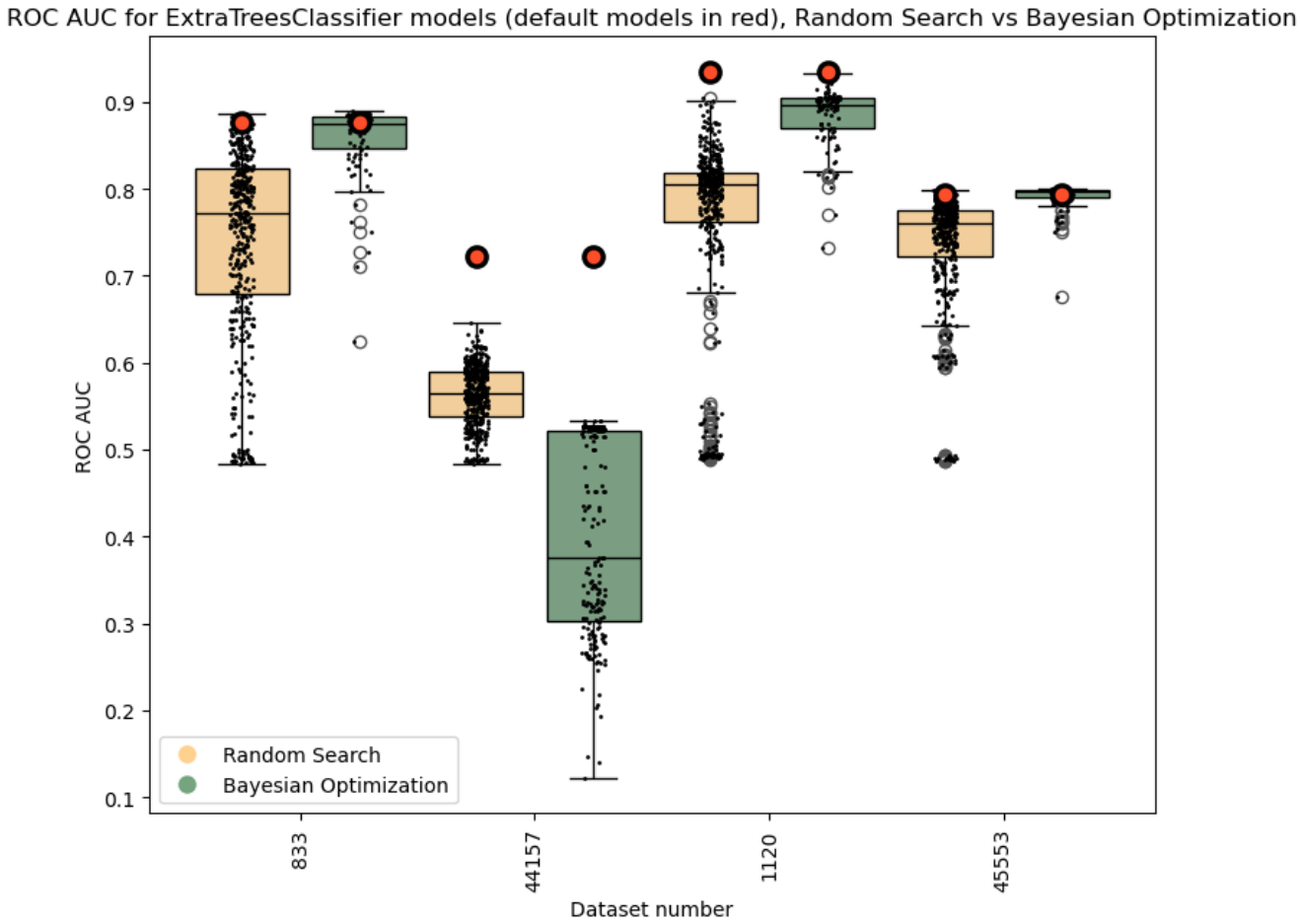
Rysunek: Porównanie wyników metryki *ROC AUC* dla metod **BayesOptimization** oraz **RandomSearch** z domyślnymi wartościami parametrów dla algorytmu **SVM** (czerwone punkty).

Metody **BayesOptimization** oraz **RandomSearch** dla **CatBoostClassifier**



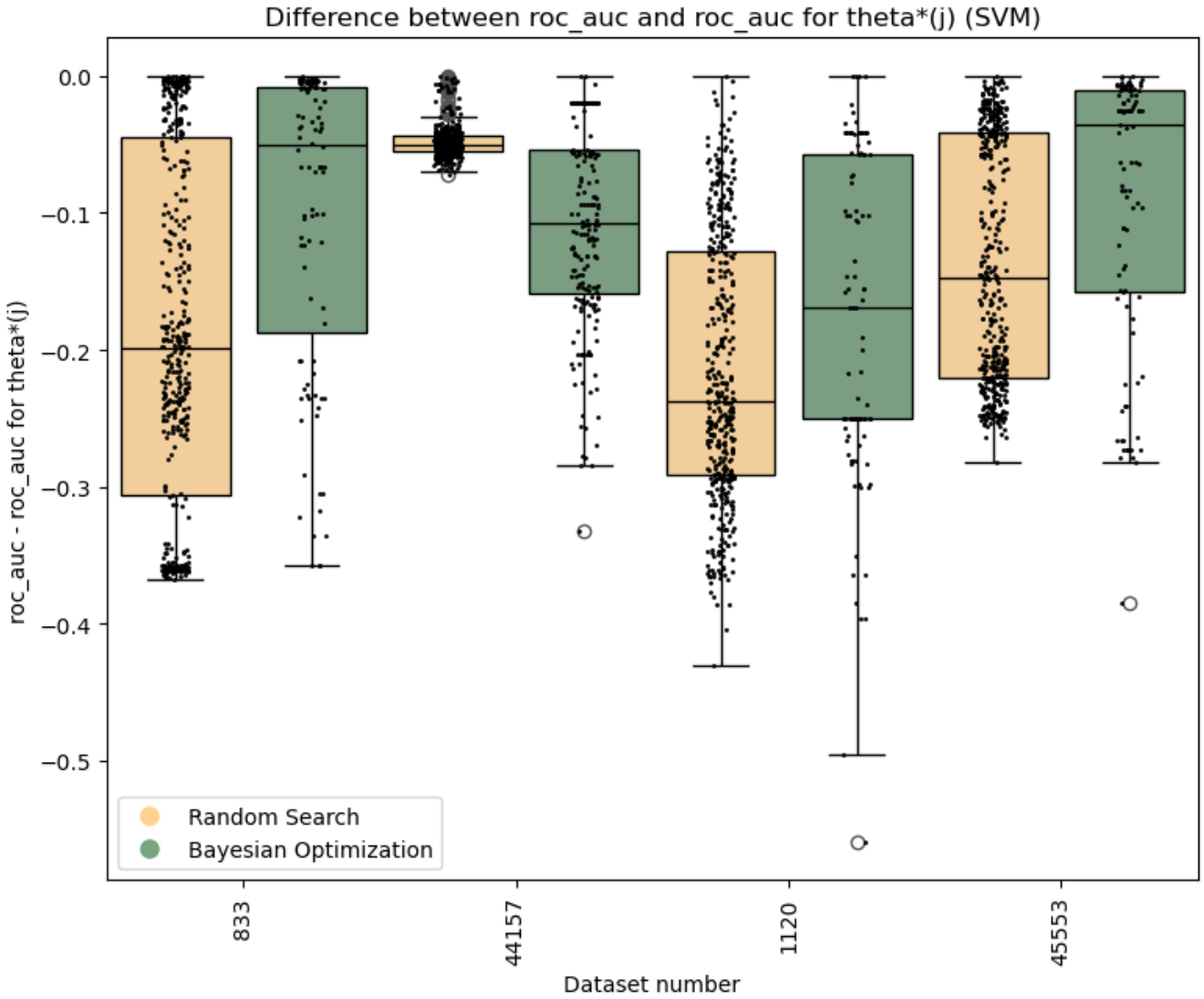
Rysunek: Porównanie wyników metryki *ROC AUC* dla metod **BayesOptimization** oraz **RandomSearch** z domyślnymi wartościami parametrów dla algorytmu **CatBoostClassifier** (czerwone punkty).

Metody **BayesOptimization** oraz **RandomSearch** dla **ExtraTreesClassifier**



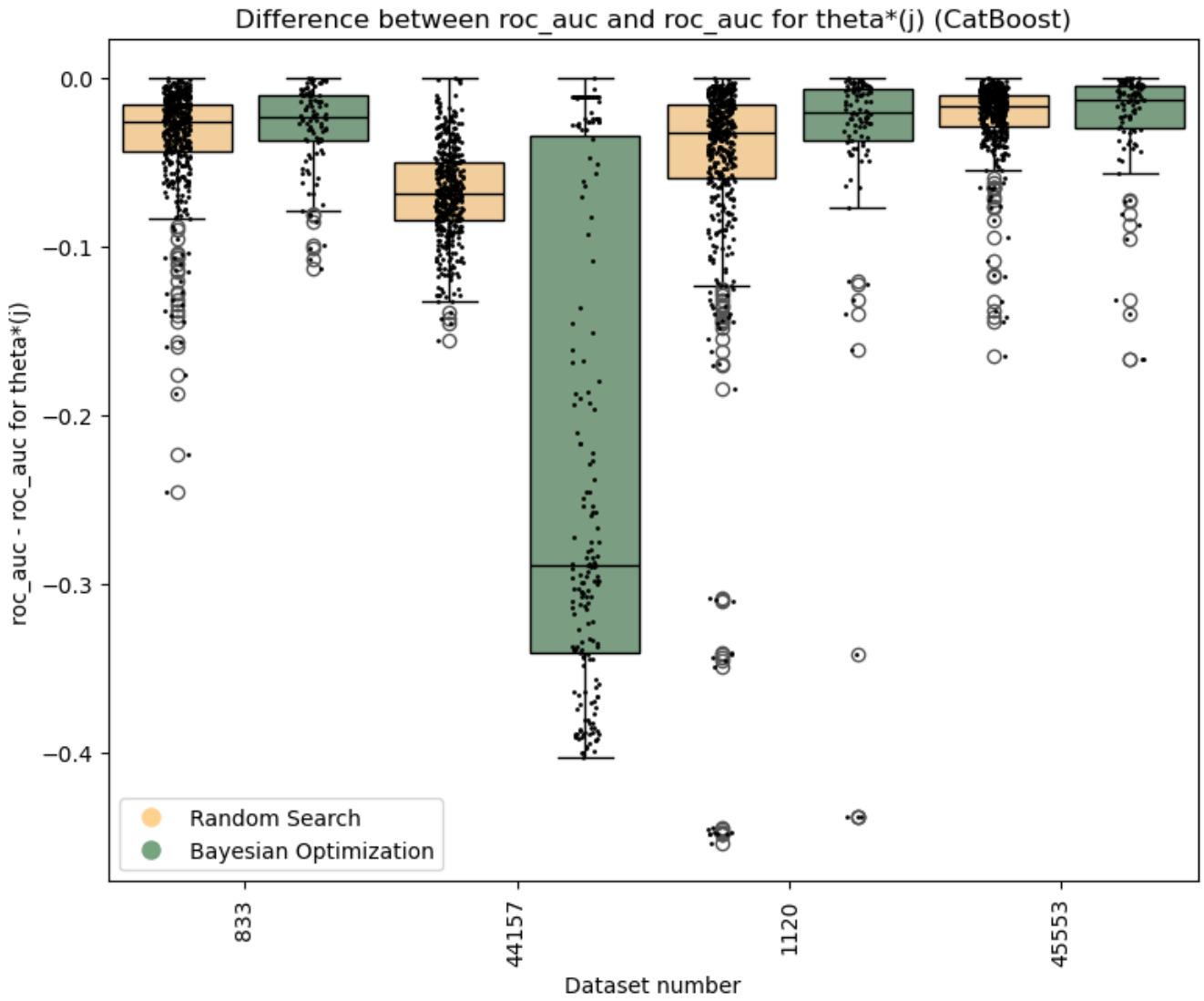
Rysunek: Porównanie wyników metryki *ROC AUC* dla metod **BayesOptimization** oraz **RandomSearch** z domyślnymi wartościami parametrów dla algorytmu **ExtraTreesClassifier** (czerwone punkty).

Metody **BayesOptimization** oraz **RandomSearch** dla **SVM**



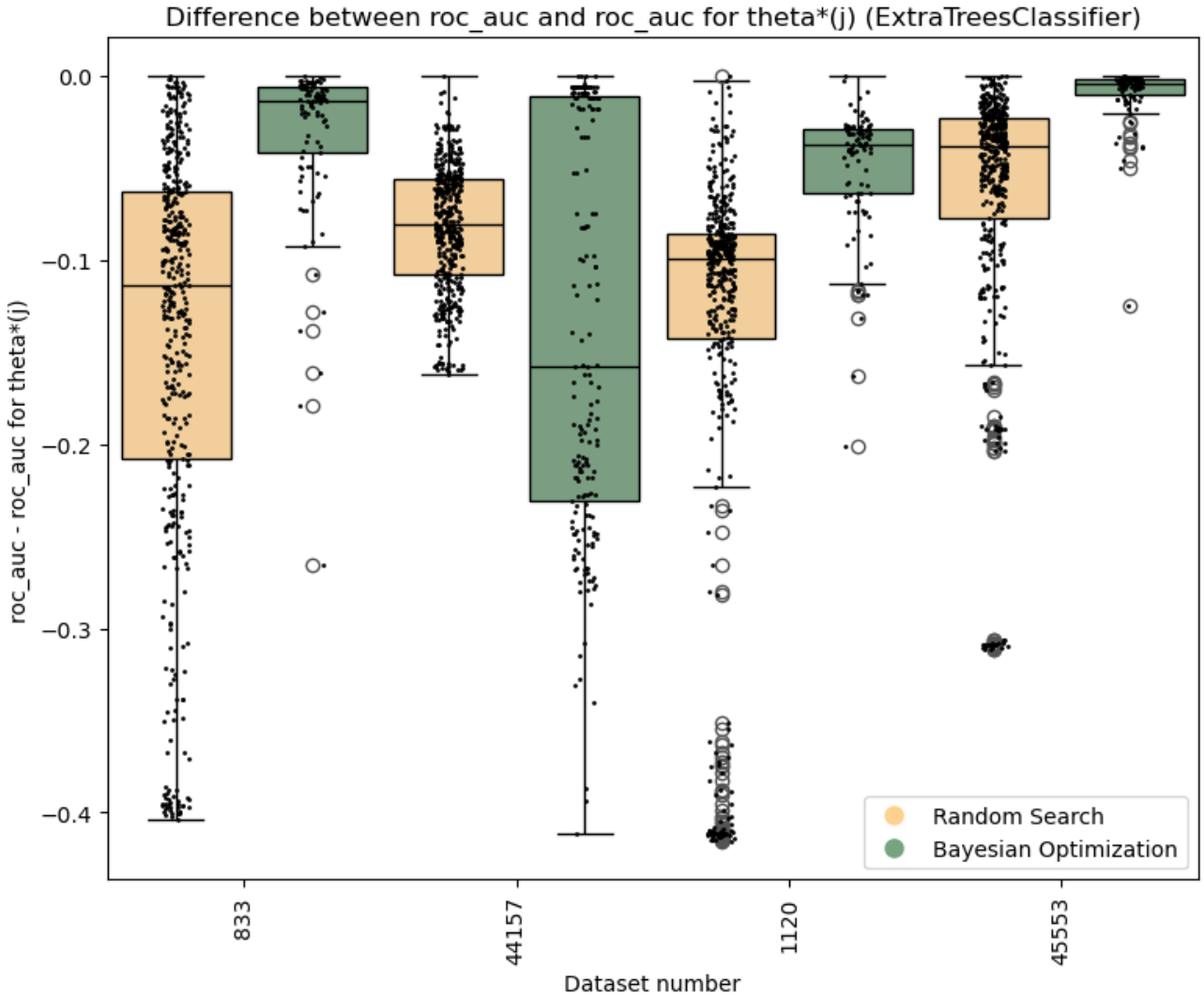
Rysunek: Różnica między *ROC AUC* metod **BayesOptimization** i **RandomSearch** dla algorytmu **SVM**.

Metody **BayesOptimization** oraz **RandomSearch** dla **CatBoostClassifier**



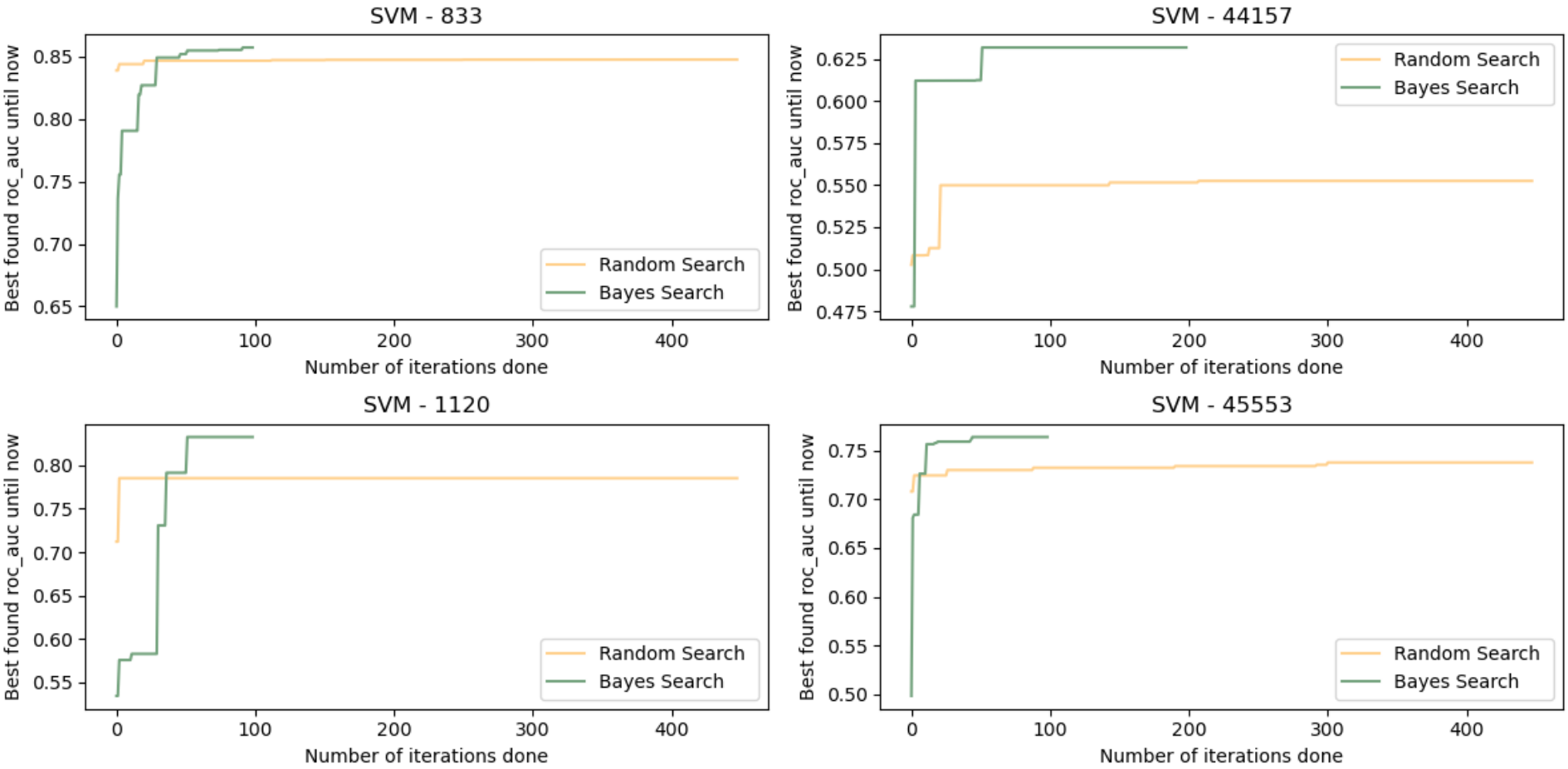
Rysunek: Różnica między *ROC AUC* metod **BayesOptimization** i **RandomSearch** dla algorytmu **CatBoostClassifier**.

Metody **BayesOptimization** oraz **RandomSearch** dla **ExtraTreesClassifier**



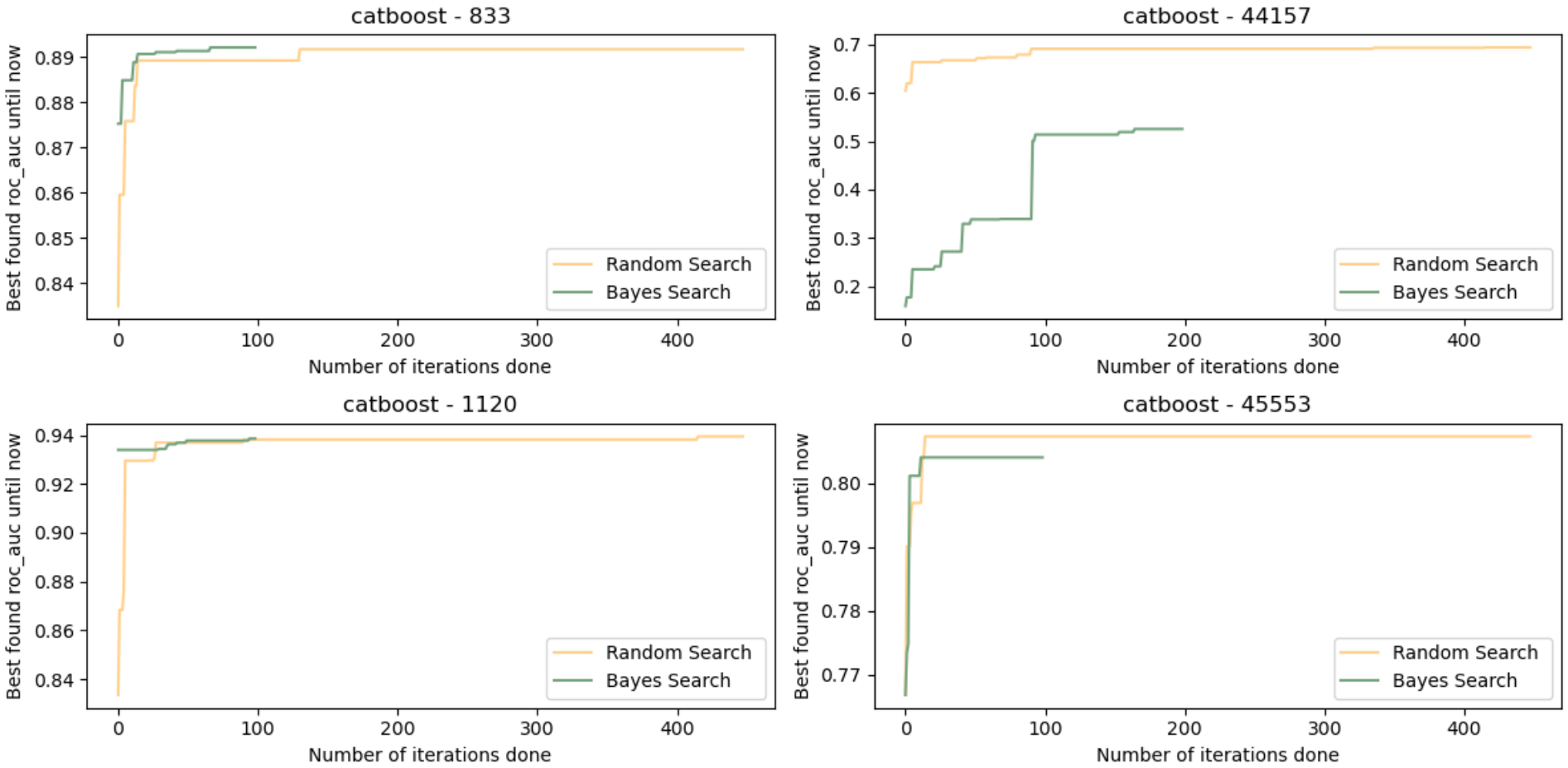
Rysunek: Różnica między *ROC AUC* metod **BayesOptimization** i **RandomSearch** dla algorytmu **ExtraTreesClassifier**.

Liczba iteracji dla **SVM**



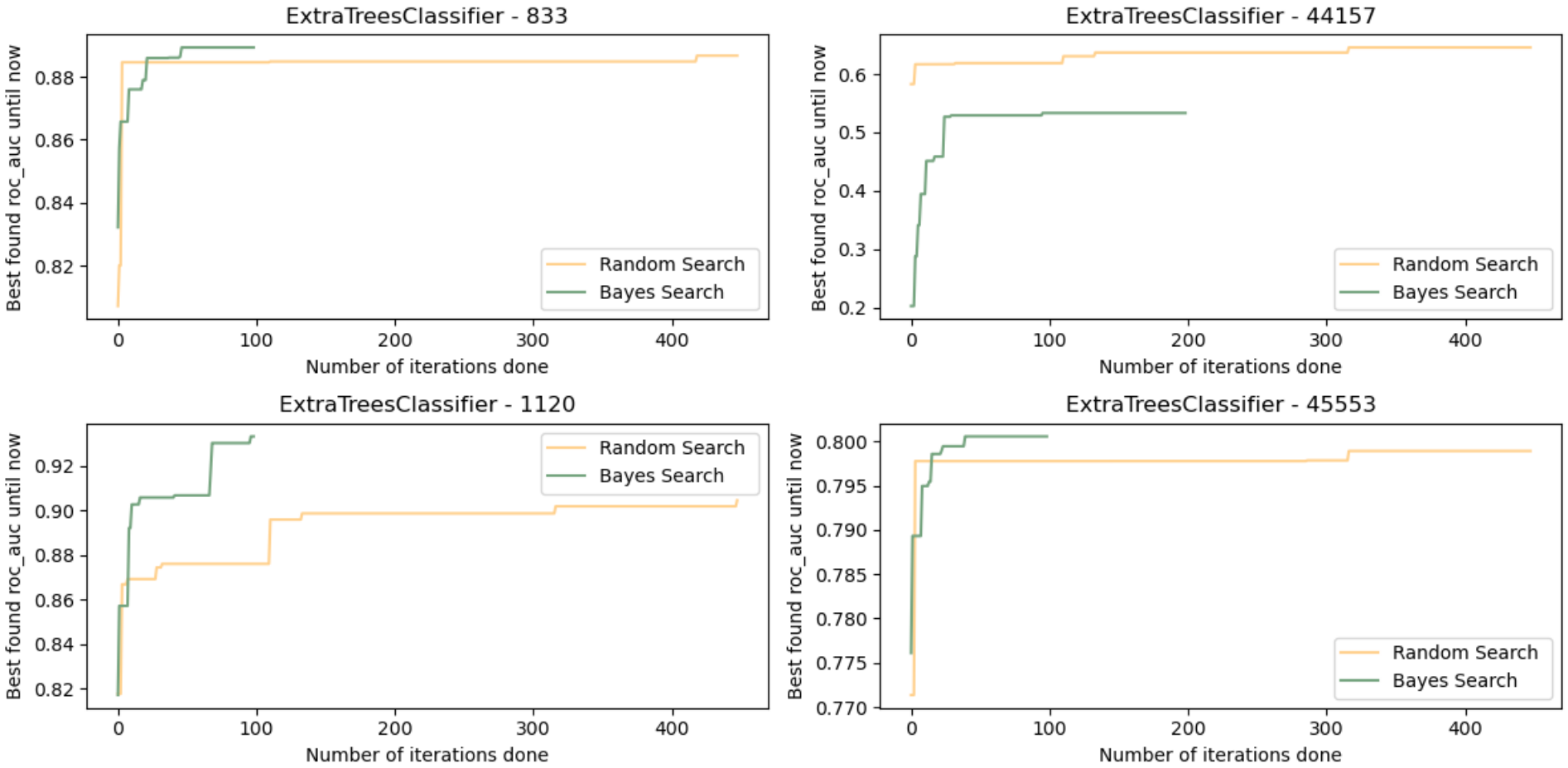
Rysunek: Liczba iteracji względem metryki *ROC AUC* metod **BayesOptimization** i **RandomSearch** dla algorytmu **SVM**.

Liczba iteracji dla **CatBoostClassifier**



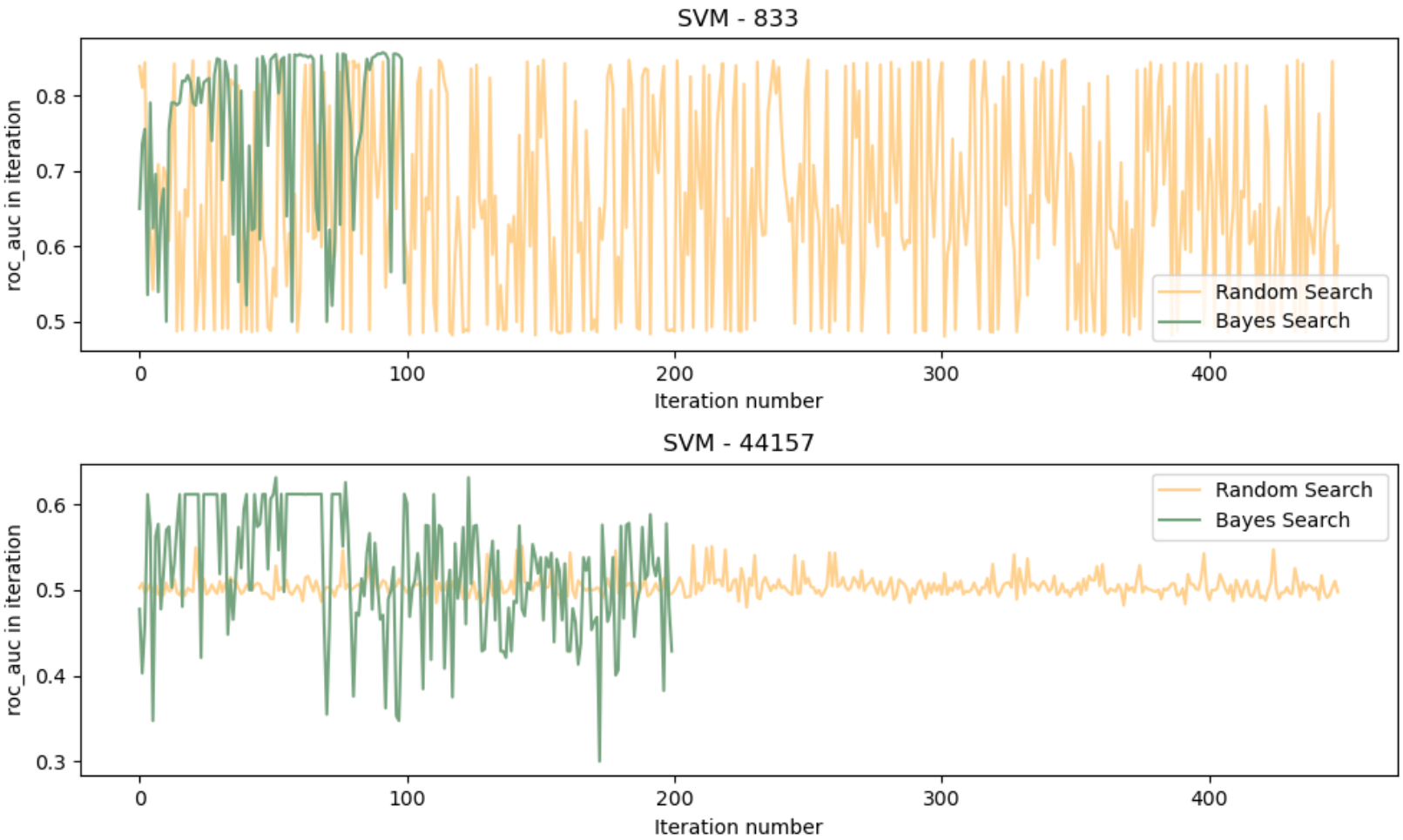
Rysunek: Liczba iteracji względem metryki *ROC AUC* metod **BayesOptimization** i **RandomSearch** dla algorytmu **CatBoostClassifier**.

Liczba iteracji dla **ExtraTreesClassifier**



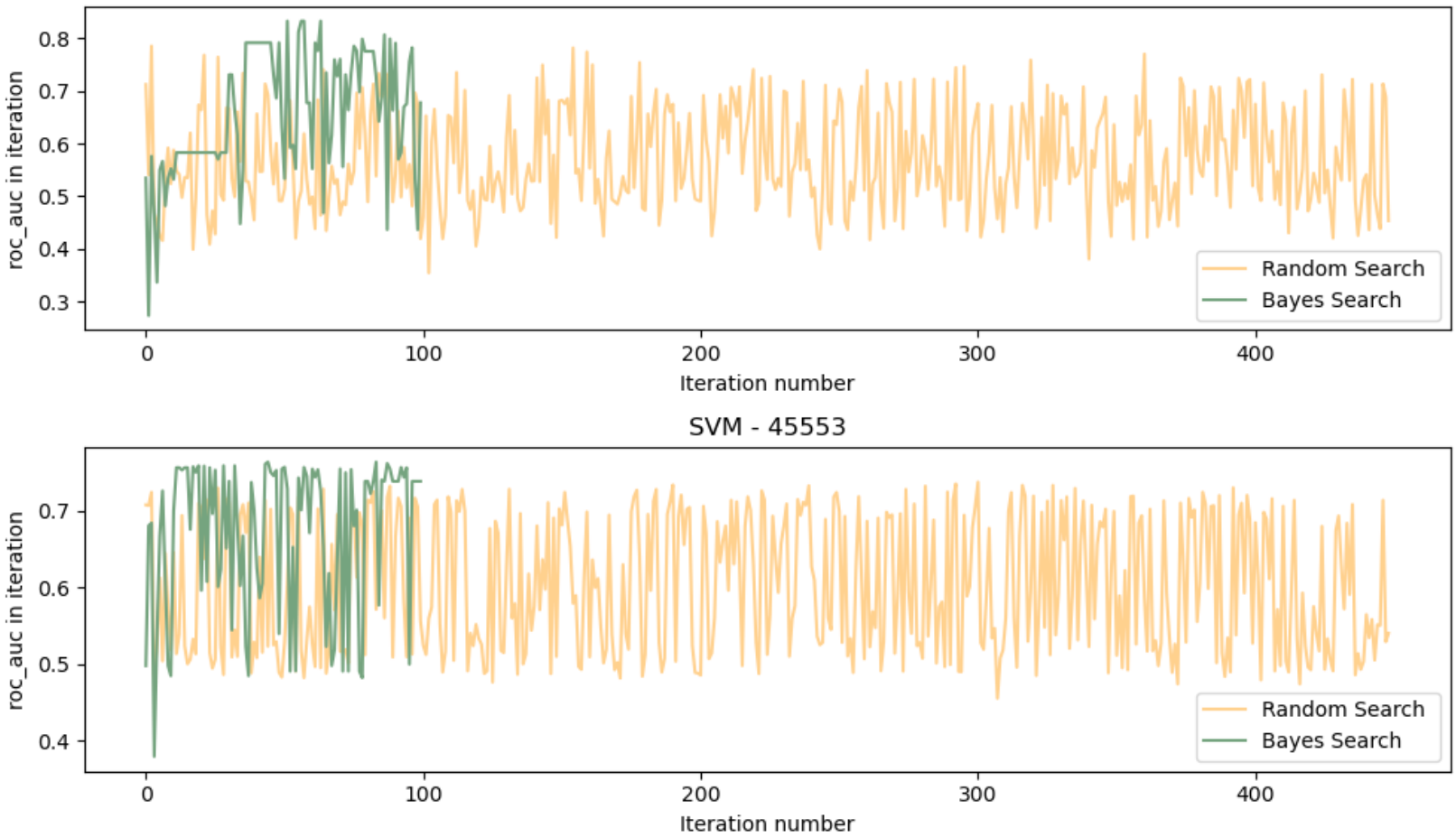
Rysunek: Liczba iteracji względem metryki *ROC AUC* metod **BayesOptimization** i **RandomSearch** dla algorytmu **ExtraTreesClassifier**.

Zmienność *ROC AUC* względem iteracji dla **SVM** - część 1



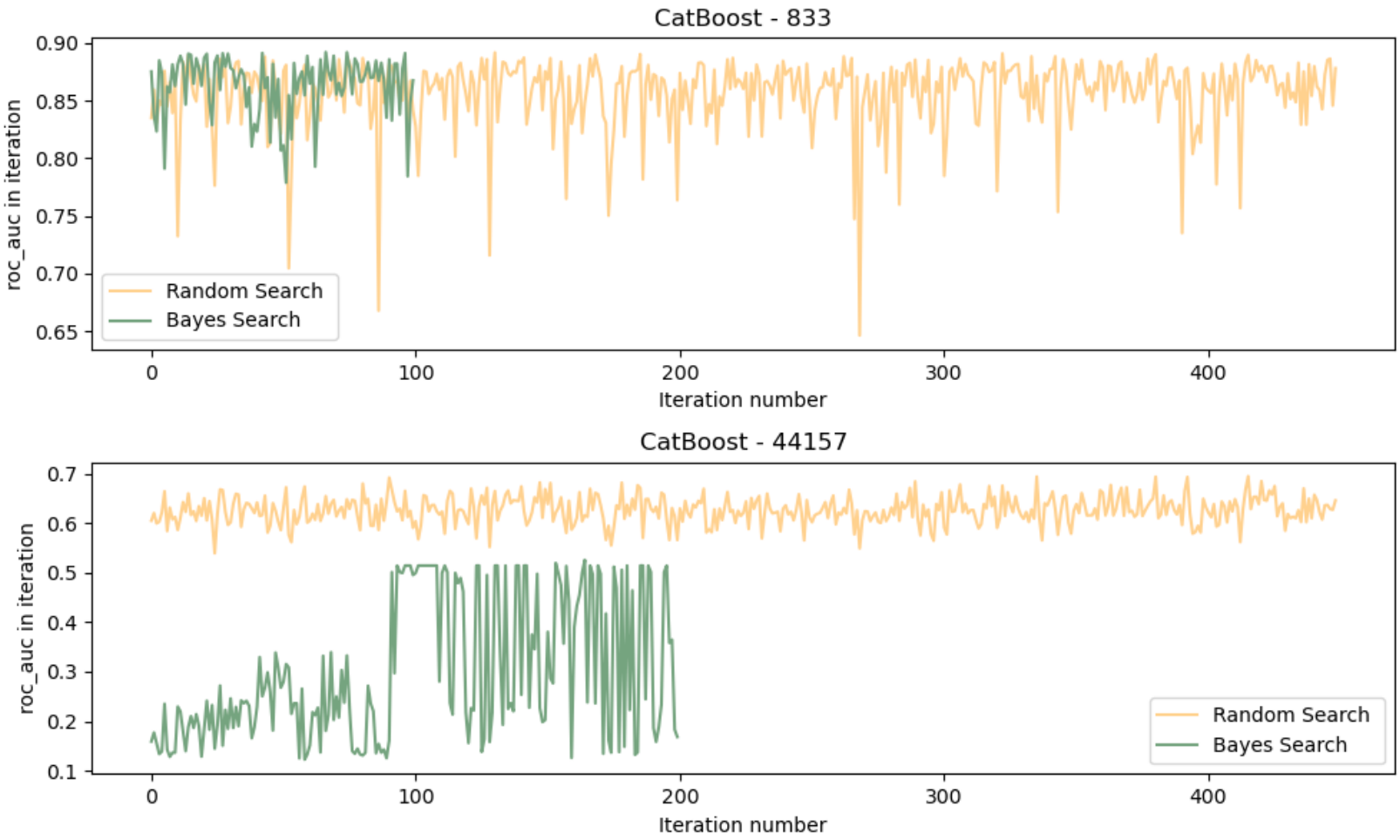
Rysunek: Zmienność metryki *ROC AUC* względem iteracji metod **BayesOptimization** i **RandomSearch** dla algorytmu **SVM**.

Zmienność *ROC AUC* względem iteracji dla **SVM** - część 2



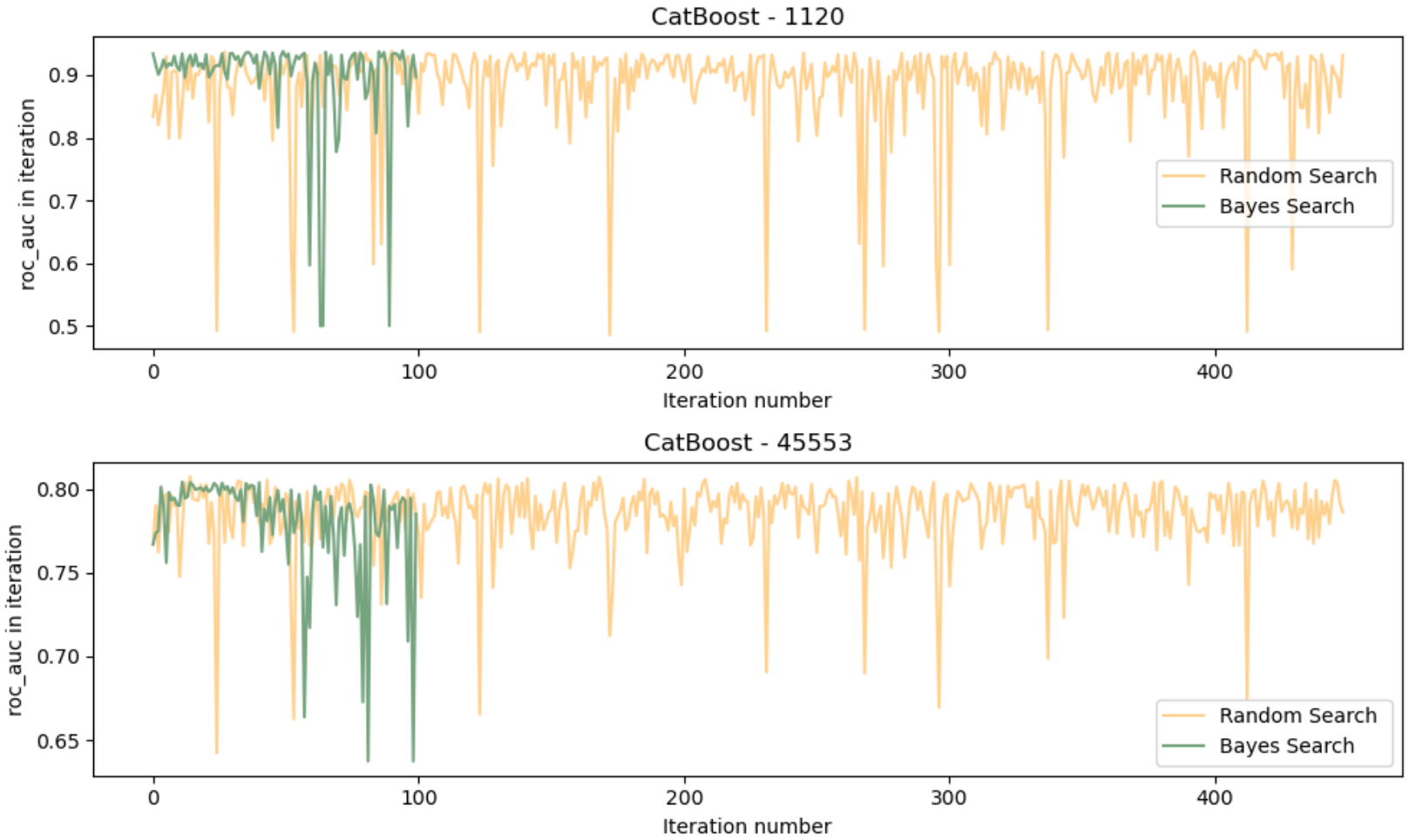
Rysunek: Zmienność metryki *ROC AUC* względem iteracji metod **BayesOptimization** i **RandomSearch** dla algorytmu **SVM**.

Zmienność *ROC AUC* względem iteracji dla **CatBoostClassifier**-część 1



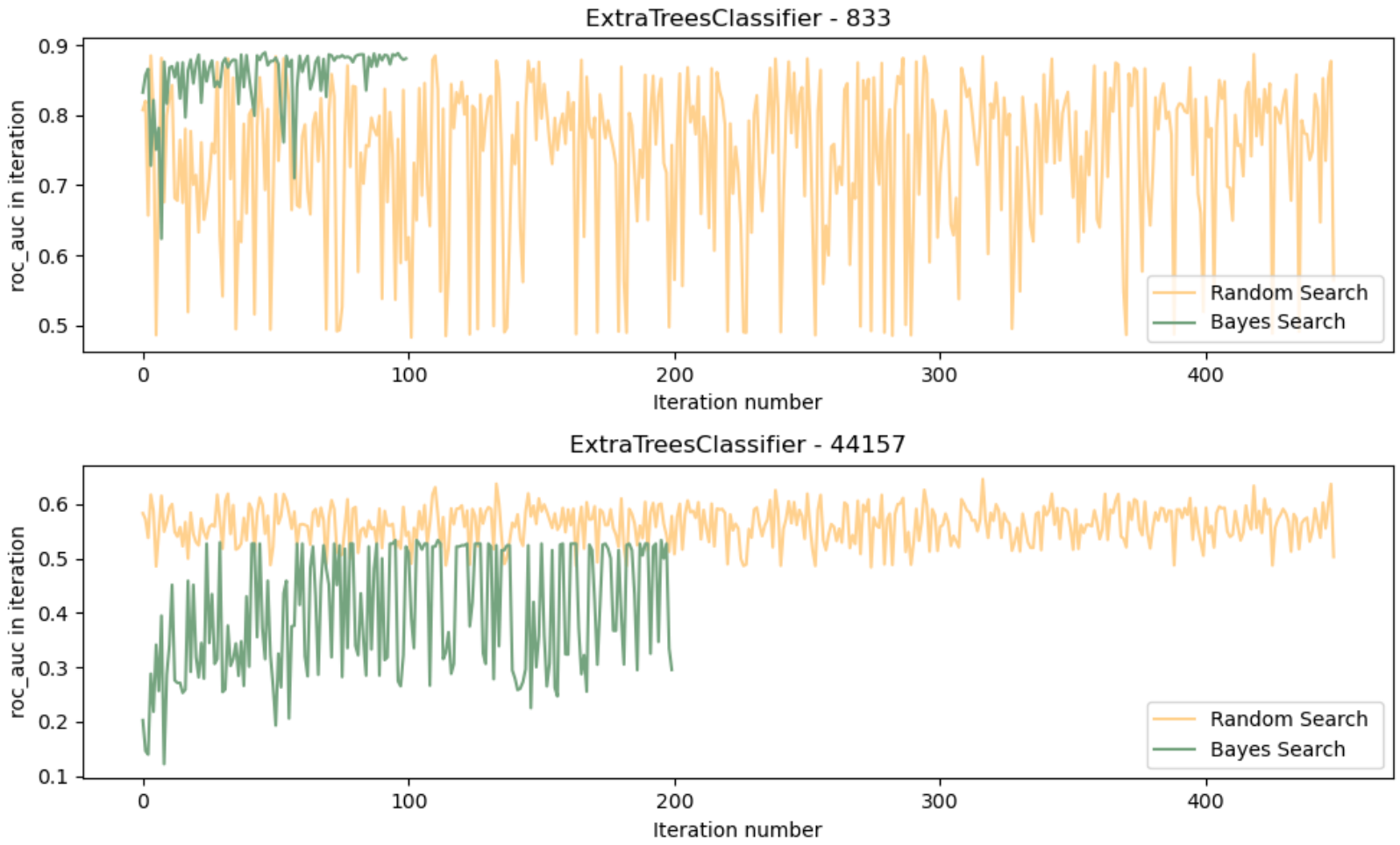
Rysunek: Zmienność metryki metryki *ROC AUC* względem iteracji metod **BayesOptimization** i **RandomSearch** dla algorytmu **CatBoostClassifier**.

Zmienność *ROC AUC* względem iteracji dla **CatBoostClassifier**-część 2



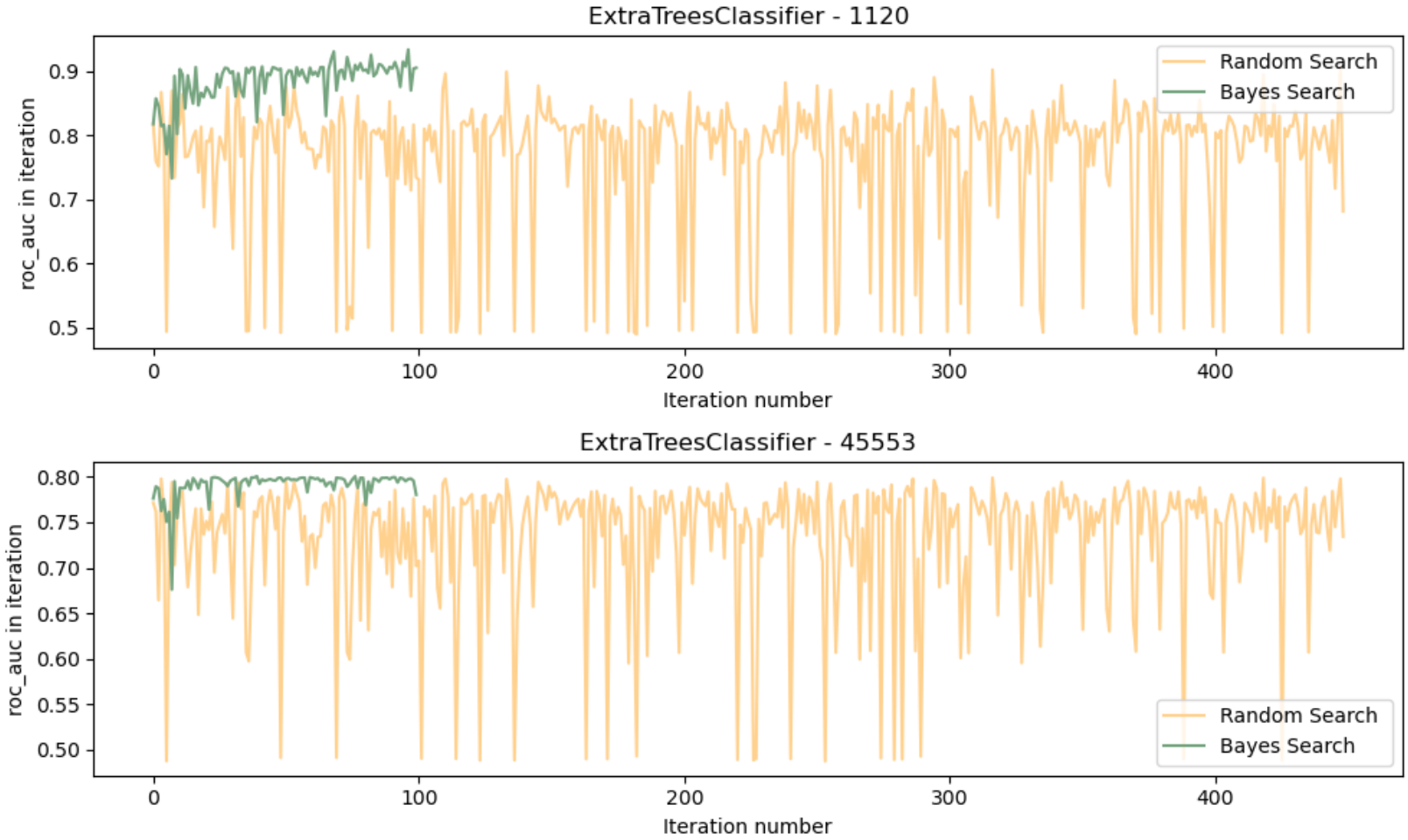
Rysunek: Zmienność metryki metryki *ROC AUC* względem iteracji metod **BayesOptimization** i **RandomSearch** dla algorytmu **CatBoostClassifier**.

Zmienność *ROC AUC* względem iteracji dla **ExtraTreesClassifier** - część 1



Rysunek: Zmienność metryki *ROC AUC* względem iteracji metod **BayesOptimization** i **RandomSearch** dla algorytmu **ExtraTreesClassifier**.

Zmienność *ROC AUC* względem iteracji dla **ExtraTreesClassifier** - część 2



Rysunek: Zmienność metryki *ROC AUC* względem iteracji metod **BayesOptimization** i **RandomSearch** dla algorytmu **ExtraTreesClassifier**.

Test Manna-Whitneya

Postać testu

Niech F_X oznacza rozkład wartości *ROC AUC* dla metody **RandomSearch**, a F_Y rozkład wartości *ROC AUC* dla metody **BayesOptimization**, wówczas test z jednostronną hipotezą alternatywną ma następującą postać

$$\begin{cases} H_0 : F_X = F_Y, \\ H_1 : X \overset{st}{<} Y. \end{cases}$$

Test Manna-Whitneya dla **SVM**

Zbiór danych	p-wartość	Wniosek
Zbiór danych 1	9.725e-14	$X \overset{st}{<} Y$
Zbiór danych 2	9.725e-14	$X \overset{st}{<} Y$
Zbiór danych 3	1.02e-12	$X \overset{st}{<} Y$
Zbiór danych 4	2.238e-13	$X \overset{st}{<} Y$

Tabela: Wyniki testu Manna-Whitneya dla algorytmu SVM.

Test Manna-Whitneya dla **CatBoostClassifier**

Zbiór danych	p-wartość	Wniosek
Zbiór danych 1	1.106e-02	$X \overset{st}{<} Y$
Zbiór danych 2	1.000	$F_X = F_Y$
Zbiór danych 3	1.121e-05	$X \overset{st}{<} Y$
Zbiór danych 4	5.301e-01	$F_X = F_Y$

Tabela: Wyniki testu Manna-Whitneya dla algorytmu CatBoost.

Test Manna-Whitneya dla **ExtraTreesClassifier**

Zbiór danych	**p-wartość**	**Wniosek**
Zbiór danych 1	4.324e-34	$X \overset{st}{<} Y$
Zbiór danych 2	1.000	$F_X = F_Y$
Zbiór danych 3	1.057e-44	$X \overset{st}{<} Y$
Zbiór danych 4	9.256e-42	$X \overset{st}{<} Y$

Tabela: Wyniki testu Manna-Whitneya dla algorytmu ExtraTreesClassifier.

Dziękujemy za uwagę!

Tunowalność wybranych algorytmów uczenia maszynowego

Katsiaryna Bokhan, Monika Jarosińska

MiNI PW

06.11.2024

Bibliografia

 P.Probst, A.Boulesteix, , Tunability: Importance of Hyperparameters of Machine Learning Algorithms, 2019.