Analiza tunowalności hiperparametrów modeli ML

Daria Bartkowiak, Alicja Przeździecka, Oliwia Wójcicka

13 listopada 2025

1 Wstęp

Celem eksperymentu było porównanie tunowalności trzech modeli uczenia maszynowego: **Logistic Regression**, **Random Forest** i **XGBoost** na czterech zestawach danych: *depression*, *diabetes*, *loan* oraz *weather*. Dla każdego modelu zastosowano dwie techniki losowego przeszukiwania przestrzeni hiperparametrów: *RandomizedSearch* oraz *BayesSearch*, przy takiej samej liczbie 50 iteracji. Analizowano wartości metryki ROC-AUC oraz stabilność wyników.

1.1 Ramki danych

W eksperymentach wykorzystano cztery zbiory danych pochodzące z platformy *Kaggle*, reprezentujące różne dziedziny problemów klasyfikacyjnych:

1. **Loan Dataset** – dane dotyczące pożyczek, służące do przewidywania, czy pożyczka zostanie spłacona (`charged_off`: 0 – spłacona, 1 – brak spłaty).
2. **Diabetes Dataset** – zbiór danych klinicznych z celem predykcyjnym określającym występowanie cukrzycy (`diabetes`: 0 – brak, 1 – obecna).
3. **Depression Dataset** – dane medyczne dotyczące występowania chorób psychicznych, z celem przewidywania historii choroby psychicznej (`History of Mental Illness`: 0 – brak, 1 – występuje).
4. **Weather Dataset** – dane meteorologiczne wykorzystywane do prognozowania, czy następnego dnia wystąpi opad deszczu (`RainTomorrow`: 0 – nie, 1 – tak).

Zbiory zostały poddane wstępnemu przetwarzaniu (czyszczenie, transformacja cech, kodowanie zmiennych kategorycznych). Dzięki zróżnicowanej tematyce – od medycyny po finanse i pogodę – dane pozwalają ocenić ogólną skuteczność oraz tunowalność badanych modeli klasyfikacyjnych.

2 Opis eksperymentu

2.1 Modele i siatki modeli

Szczegółowe definicje wszystkich przestrzeni przeszukiwania hiperparametrów dla metod **RandomizedSearch** oraz **BayesSearch** znajdują się w osobnym załączniku:

Załącznik 1: siatki.pdf

3 Wyznaczenie najlepszych zestawów hiperparametrów

W ramach eksperymentu przeprowadzono dostrajanie modeli dla wszystkich czterech zbiorów danych: *depression*, *diabetes*, *loan* oraz *weather*. Dla każdego z modeli — **Logistic Regression**, **Random Forest** oraz **XGBoost** — znaleziono **najlepszy zestaw hiperparametrów** osobno dla metod *RandomizedSearch* i *BayesSearch*. Każde przeszukiwanie wykonywano przez **50 iteracji**, co pozwoliło uzyskać stabilne wyniki optymalizacji.

Dodatkowo zbadano **wpływ liczby iteracji** na stabilność wyniku i szybkość zbieżności algorytmu optymalizacji. Wykresy ilustrujące tę zależność zostały umieszczone w załączniku:

Załącznik 2: liczba_iteracji.pdf

Analiza tych wykresów wykazała, że w większości przypadków — niezależnie od modelu, ramki danych i zastosowanej metody przeszukiwania — **najlepszy wynik został osiągnięty znacznie wcześniej niż w 50. iteracji**, a kolejne próby nie prowadziły do dalszej poprawy jakości modelu. Oznacza to, że proces optymalizacji zbiegał szybko i osiągał stabilne maksimum metryki ROC AUC.

Na podstawie wyników uzyskanych dla wszystkich czterech zbiorów danych wyznaczono również **uniwersalny zestaw hiperparametrów** dla każdego modelu. W tym celu przeanalizowano rezultaty uzyskane dla poszczególnych ramek danych i na ich podstawie wybrano zestaw konfiguracji, który dawał **najlepszy średni wynik spośród wszystkich zestawów hiperparametrów**. Tym samym domyślny zestaw hiperparametrów reprezentuje uśrednioną, najbardziej efektywną konfigurację spośród analizowanych przypadków.

Wszystkie **zestawy najlepszych i domyślnych hiperparametrów** zostały zapisane w pliku Pythonowym `tunning_all_models.ipynb`.

4 Analiza tunowalności modeli

4.1 Metodyka

W celu oceny **tunowalności** modeli zastosowano podejście oparte na analizie różnic wyników jakości predykcji (ROC AUC) pomiędzy zestawami dobranych hiperparametrów a zestawem domyślnym. Dla każdego modelu oraz każdej ramki danych (zestawu cech) obliczano różnicę pomiędzy wartością ROC AUC uzyskaną dla danego zestawu hiperparametrów a wartością osiągniętą przy zastosowaniu zestawu domyślnego:

$$\Delta AUC_{m,r,p} = AUC_{m,r,p} - AUC_{m,r,\text{default}},$$

gdzie m oznacza model, r – ramkę danych, a p – zestaw hiperparametrów.

Tak wyznaczone różnice ΔAUC zebrano w jedną ramkę danych i posłużyły one do dalszej analizy. Wyniki zaprezentowano w formie wykresów typu *heatmap* oraz *boxplot*, które przedstawiają

rozkład wartości ΔAUC dla każdego modelu oraz metody strojenia hiperparametrów (*Random Search* oraz *Bayesian Search*). Wykresy pudełkowe umożliwiają ocenę **rozrzutu wyników**, **median**, **wartości skrajnych** oraz **obecności obserwacji odstających**, co pozwala oszacować, jak silnie wynik modelu zależy od zastosowanego zestawu hiperparametrów. Wysoka zmienność (szerokie pudełka i długie wąsy) wskazuje na dużą **tunowalność** modelu, natomiast niewielki rozrzut sugeruje, że model jest stabilny i mniej podatny na zmianę parametrów.

4.2 Wizualizacja wyników

Wykresy pudełkowe oraz heatmap'y dla poszczególnych modeli oraz metody strojenia hiperparametró zostały przedstawione w załączniku.

Załącznik 3: tunowalnosc.pdf

4.3 Wnioski

- **XGBoostClassifier** – model o najwyższej tunowalności. Mediana ΔAUC jest dodatnia, a rozrzut szeroki, co wskazuje, że tuning hiperparametrów zazwyczaj poprawia wynik modelu. XGBoost silnie reaguje na zmiany parametrów, co pozwala uzyskać istotne zyski w jakości predykcji przy odpowiednim dostrojeniu.
- **RandomForestClassifier** – tunowalność umiarkowana i niestabilna. Duży rozrzut wartości ΔAUC oraz mediana bliska zera świadczą o tym, że efektywność strojenia zależy od konkretnego zestawu danych, a tuning nie zawsze prowadzi do poprawy wyniku.
- **LogisticRegression** – model o najniższej tunowalności. Wartości ΔAUC skupiają się w pobliżu zera, a rozrzut jest niewielki, co oznacza, że zmiana hiperparametrów ma ograniczony wpływ na skuteczność klasyfikacji. Model jest mało podatny na tuning.

4.4 Wpływ techniki losowania punktów na ocenę tunowalności

Porównanie wyników uzyskanych dla metod *Random Search* oraz *Bayesian Search* (Załącznik 3.) nie wskazuje na istotne różnice w rozkładzie wartości ΔAUC pomiędzy tymi podejściami. Dla wszystkich analizowanych modeli mediany oraz rozrzuty wyników są zbliżone, co sugeruje, że sposób próbkowania punktów w przestrzeni hiperparametrów nie wpływa znacząco na końcową ocenę tunowalności modeli.

W szczególności, dla modeli *XGBoostClassifier* oraz *RandomForestClassifier* obie metody generują podobne rozkłady różnic AUC, a niewielkie odchylenia można uznać za wynik losowości procesu strojenia, a nie systematycznego błędu (biasu). Dla modelu *LogisticRegression* różnice między metodami są minimalne, co potwierdza jego niską wrażliwość na tuning hiperparametrów.

Podsumowując, nie zaobserwowano efektu *sampling bias* – technika losowania punktów nie zmienia istotnie wniosków dotyczących tunowalności algorytmów. Oznacza to, że oba podejścia prowadzą do spójnych ocen wpływu hiperparametrów na jakość modeli.

5 Analiza stabilności wyników tuningu w zależności od podziału danych

Dodatkowo przeanalizowano stabilność wyników modeli w zależności od podziału danych treningowych i testowych (*splitów*) oraz zastosowanej metody strojenia hiperparametrów. W tym celu porównano wartości metryki ROC AUC uzyskane w kolejnych podziałach dla modeli dostrojonych metodami *Random Search* oraz *Bayesian Search*. Wyniki zostały przedstawione na wykresach w załączniku:

Załącznik 4: stabilnosc.pdf

Analiza wykazała, że wyniki modeli pozostają generalnie stabilne pomiędzy kolejnymi podziałami danych, co świadczy o dobrej powtarzalności procesu trenowania.

- **RandomForestClassifier** okazał się modelem o największej zmienności wyników (choć i tak nie bardzo dużej). Oznacza to, że stabilność lasu losowego lekko zależy od charakterystyki danych wejściowych.
- **XGBoostClassifier** oraz **LogisticRegression** wykazały wysoką stabilność, z niewielkimi odchyleniami pomiędzy splitami. Dla tych modeli zarówno metoda losowa, jak i bayesowska dawały bardzo zbliżone wyniki w kolejnych podziałach danych.

6 Ewaluacja modeli na zbiorze testowym

W celu oceny jakości modeli po dostrojeniu hiperparametrów przeprowadzono testowanie na niezależnym zbiorze testowym. Dla każdego modelu (**Logistic Regression**, **Random Forest**, **XGBoost**) oraz każdego zbioru danych (*depression*, *diabetes*, *loan*, *weather*) obliczono wartość metryki **ROC-AUC**.

Wyniki uzyskane dla najlepszych oraz domyślnych zestawów hiperparametrów (metoda **RandomizedSearch**) przedstawiono w załączniku:

Załącznik 5: wyniki.pdf

6.1 Wnioski

Wyniki testowe potwierdzają wysoką skuteczność i dobrą generalizację dostrojonych modeli.

- Wszystkie modele osiągnęły **wysokie wartości ROC-AUC** (Zał. 5), co potwierdza skuteczność tuningu.
- **XGBoost** uzyskał najwyższe wyniki, zwłaszcza dla zbiorów *Diabetes* i *Loan* (AUC > 0.97).
- **Random Forest** był stabilny i zbliżony do XGBoost, a **Logistic Regression** uzyskała konkurencyjne rezultaty mimo prostszej struktury.
- Niższe wyniki dla *Depression* wynikają z trudniejszej charakterystyki danych.