Wydział Matematyki
i Nauk Informacyjnych

POLITECHNIKA WARSZAWSKA

AutoML raport 2

Metody klasyfikacji

Devid Khodak 317253
Kamil Kubiak 313293

16 stycznia 2024

Spis treści

1 Wstęp

Celem sprawozdania jest ocena oraz porównanie modeli ręcznych (dobór rodzaju modelu i hiperparametrów zależy od nas) z modelami AutoML'owymi, a także sprawdzenie jaki wpływ na końcową jakość modelu ma preprocessing.

2 Zbiór danych

Korzystamy ze sztucznie wygenerowanego zbioru danych $artificial_train$, który oryginalnie ma 500 zmiennych.

2.1 Preprocessing

Najpierw próbowaliśmy pozbyć się zmiennych o zerowej wariancji, okazało się jednak, że takie nie istnieją, zatem ten etap nic nie zmienił w zbiorze danych.
Następnie, usunięte zostały wszystkie zmienne o wysokiej korelacji, czyli większej od 0.9 - ten krok pozwolił nam wyeliminować 10 zmiennych.
Ostatnim krokiem preprocessingu było Random Forest feature selection, które już znacząco zredukowało wymiar, bo zostało wyłącznie 9 zmiennych.

3 Modele ręczne

Do wyznaczenia najlepszej konfiguracji hiperparametrów zastosowano metodę Random Search. Modele bez preprocessingu (na 500 zmiennych) były trenowane dla 5 iteracji przy 5-krotnej kroswalidacji. Natomiast modele po redukcji wymiaru (na 9 zmiennych) były trenowane dla 50 iteracji przy 5-krotnej kroswalidacji.

3.1 XGBoost

Rozważana była następująca siatka hiperparametrów:

- gamma: [0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 5]
- learning_rate: [0.01, 0.03, 0.06, 0.1, 0.15, 0.2, 0.25, 0.3]
- max_depth: [3, 5, 6, 7, 8, 9, 10]
- n_estimators: [50, 75, 100, 125, 150, 175, 200]
- reg_alpha: [0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 200]
- reg_lambda: [0, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4, 200]

W tabeli 1 zostały umieszczone wyniki Balanced accuracy dla 5-krotnej kroswalidacji wraz z konfiguracją dla której uzyskano ten wynik.

Hiperparametry	**Model bez preprocessingu**	**Model po redukcji wymiaru**
gamma	0.1	0.2
learning_rate	0.2	0.06
max_depth	6	10
n_estimators	75	125
reg_alpha	3.2	1.6
reg_lambda	0.2	0.2
Balanced accuracy	0.79	0.875

Tabela 1: Zestawy hiperparametrów, dla których osiagane były najwyższe wyniki dla algorytmu XGBoost

3.2 Random forest

Rozważana była następująca siatka hiperparametrów:

- n_estimators: [50, 75, 100, 125, 150, 175, 200]
- max_features: ['log2', 'sqrt', None]
- max_depth: [5, 10, 15, 20, 25, None]
- min_samples_split: [2, 5, 10]
- min_samples_leaf: [1, 2, 4]
- bootstrap: [True, False]

W tabeli 2 zostały umieszczone wyniki Balanced accuracy dla 5-krotnej kroswalidacji wraz z konfiguracją dla której uzyskano ten wynik.

Hiperparametry	**Model bez preprocessingu**	**Model po redukcji wymiaru**
n_estimators	125	100
min_samples_split	10	5
min_samples_leaf	4	4
max_features	None	sqrt
max_depth	None	15
bootstrap	True	False
Balanced accuracy	0.7955	0.8815000000000002

Tabela 2: Zestawy hiperparametrów, dla których osiagane były najwyższe wyniki dla algorytmu Random Forest

3.3 Najlepszy model

Najlepszy model uzyskano dla algorytmu Random Forest na zbiorze danych po redukcji wymiaru. Balanced accuracy na całym zbiorze treningowym dla 5-krotnej kroswalidacji wyniosło 0.8815.

4 Modele AutoML

Automatyczne modele, które wybraliśmy to AutoSklearn, AutoSklearn2 oraz AutoGluon. Obu modelom AutoSklearn daliśmy po 10 minut na trenowanie. a wybraną metryką było *balanced accuracy*. Dla każdego z modeli sprawdziliśmy zestawy danych bez preprocessingu oraz z nim. Poniżej znajdują się wyniki:

Balanced accuracy	**Model bez preprocessingu**	**Model po redukcji wymiaru**
AutoSklearn	0.8475	0.8679
AutoSklearn2	0.8434	0.8654
AutoGluon	0.8334	0.8718

Tabela 3: Wyniki balanced accuracy modeli AutoML

Możemy zauważyć, że preprocessing poprawia wyniki chociaż raczej nieznacznie, warto też dodać, że przy eksperymentach z czasem, który dawaliśmy modelom AutoSklearn i AutoSklearn2 nie zauważyliśmy wielkich różnic pomiędzy 2 minutami, a 10 minutami, w pierwszym przypadku udało się już osiągnąć *balancedaccuracy* przekraczające 0.8 bez względu na obecność preprocessingu.
Najlepszym z modeli okazał się AutoGluon w przypadku z redukcją wymiaru, natomiast bez preprocessingu najlepiej spisał się AutoSklearn. Różnice między modelami automatycznymi są jednak w obu przypadkach bardzo małe i ciężko na podstawie tych wyników wyciągać daleko idące wnioski, o przewadze jednego nad drugim, jedyne co możemy stwierdzić to podobną jakość każdego z modeli.

4.1 Najlepszy model

Niewielką różnicą najlepszym modelem okazał się AutoGluon z redukcją wymiaru.

5 Podsumowanie - porównanie dwóch rozwiązań

Widzimy, że w modelach bez preprocessingu lepiej spisują się wszystkie modele AutoML, natomiast redukcja wymiarów bardzo pomaga modelom ręcznym, które po takiej modyfikacji danych okazują się sprawować lepiej niż każdy z modeli automatycznych, choć różnice są naprawdę niewielkie.