Wstęp do uczenia maszynowego
Praca domowa 1

Franciszek Saliński

24 października 2024

1 Opis zadania

Moim zadaniem było sprawdzenie jak hiperparametry modelu drzewa decyzyjnego oraz rozmiar zbioru uczącego wpływają na jakość predykcji modelu, a także wybranie jednego najlepszego moim zdaniem modelu i ocena jego działania.

2 Przygotowanie do pracy z danymi

W pierwszej kolejności sprawdziłem typy zmiennych oraz czy dane zawierają braki. Pozbyłem się 3 rekordów, dla których zmienna celu była nieznana. Przyjrzałem się także rozkładom zmiennych. Zmienna celu jest zbalansowana, co było świetną informacją przed modelowaniem. Zauważyłem, że niektóre zmienne bardzo dobrze rozróżniają klasę 0 od 1, a inne nie. Najprawdopodobniej te pierwsze będą kluczowe w predykcji dla naszych modeli.

3 Eksperyment

Pierwszym elementem zadania było przeprowadzenie eksperymentu ukazującego miarę AUC drzewa decyzyjnego nauczonego na zbiorze treningowym i testowym. W tym celu dla każdego zestawu hiperparametrów wykorzystałem 5-krotną kroswalidację na danych treningowych i liczyłem średnią miarę AUC, a także trenowałem drzewo na całym zbiorze treningowym i liczyłem AUC dla predykcji na zbiorze testowym.

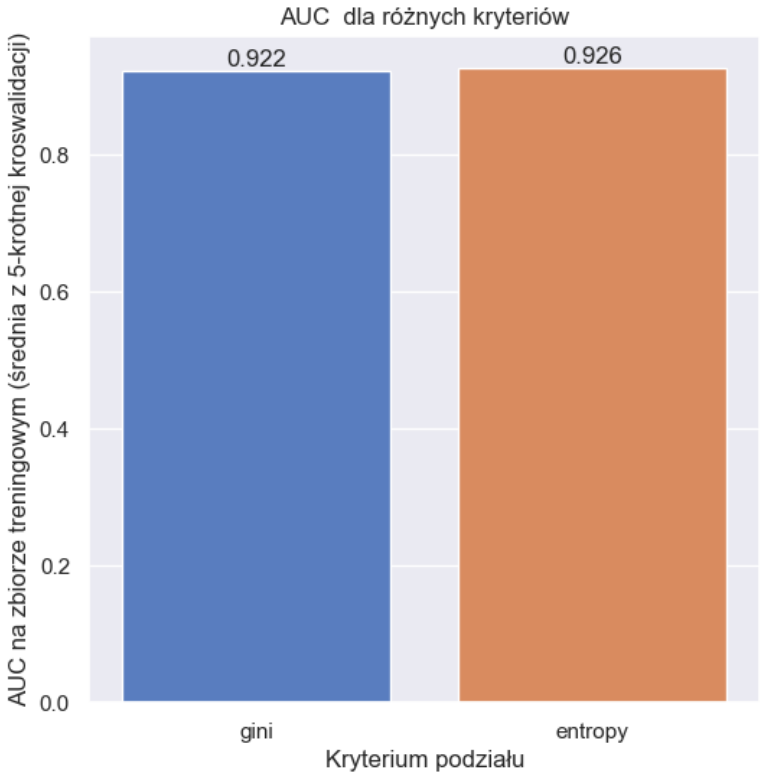
3.1 Siatka parametów

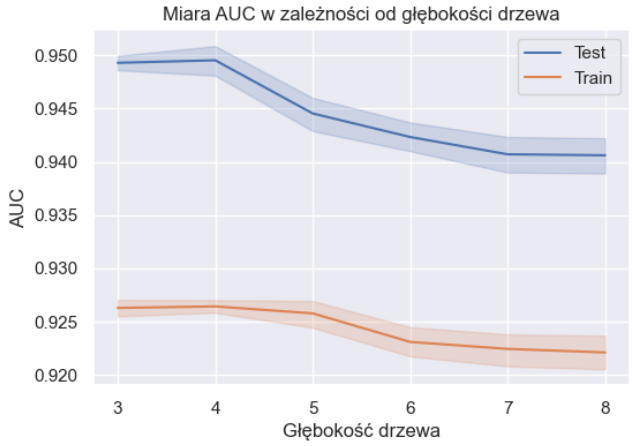
Na podstawie dokumentacji, przykładów i posiadanej wiedzy postanowiłem wybrać takie wartości parametrów do sprawdzania:

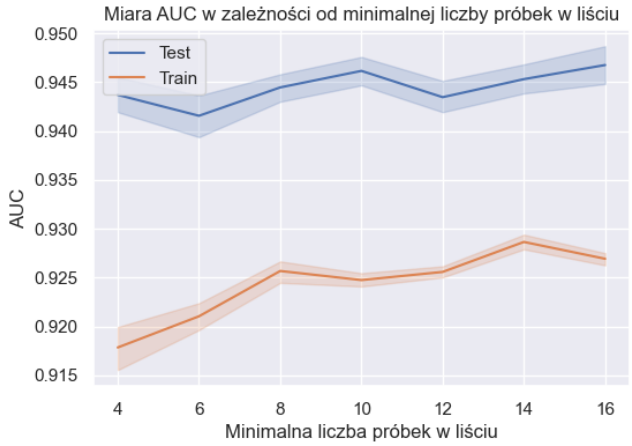
- kryterium podziału (criterion $\in \{$"*gini*", "*entropy*"$\}$)
- głebokość drzewa (max_depth $\in \{3, 4, 5, 6, 7, 8\}$)
- minimalna liczba próbek w liściu (min_samples_leaf $\in \{4, 6, 8, 10, 12, 14, 16\}$)
- maksymalna liczba liści (max_leaf_nodes $\in \{8, 12, 16, 20, 24, 28\}$)

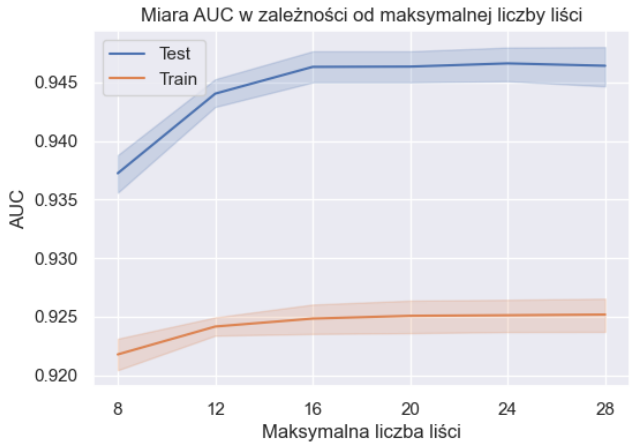
3.2 Wyniki eksperymentu

Trenując model ze wszystkimi możliwymi kombinacjami hiperparametrów, po odpowiednich agregacjach otrzymałem następujące wyniki:




Miara AUC w zależności od głębokości drzewa
Test
Train
AUC
0.950
0.945
0.940
0.935
0.930
0.925
0.920
3
4
5
6
7
8
Głębokość drzewa


Miara AUC w zależności od minimalnej liczby próbek w liściu
Test
Train
AUC
0.950
0.945
0.940
0.935
0.930
0.925
0.920
0.915
4
6
8
10
12
14
16
Minimalna liczba próbek w liściu


Miara AUC w zależności od maksymalnej liczby liści
Test
Train
AUC
0.945
0.940
0.935
0.930
0.925
0.920
8
12
16
20
24
28
Maksymalna liczba liści

3.3 Wnioski

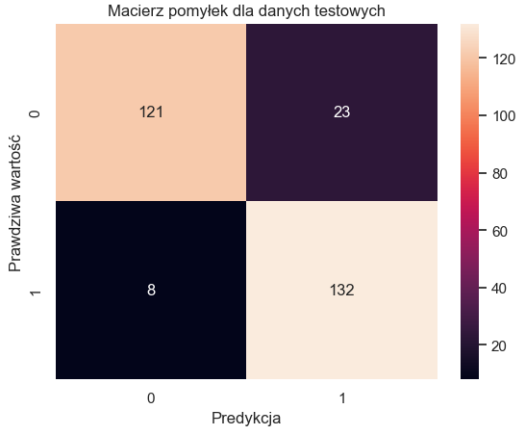
- Kryterium podziału stosowane w budowie drzew nie miało żadnego znaczącego wpływu na wartość AUC
- Dla wszystkich wartości każdego z parametrów uzyskaliśmy lepszy wynik dla danych testowych niż treningowych, co może być bardzo zaskakujące. Może to wynikać z metodyki przeprowadzania eksperymentu (dla danych treningowych liczyliśmy średnią z 5 foldów kroswalidacji, a dla danych testowych metrykę dla modelu zbudowanego na całym zbiorze treningowym), ale może być także wywołane przypadkiem (stosowane $random\_state$).
- Lepsze wyniki na zbiorze testowym niż treningowym sugerują, że raczej nie mieliśmy do czynienia z przetrenowywaniem się drzew
- Płytsze drzewa (głębokość od 3 do 5) dawały co do zasady lepsze wyniki
- Maksymalna liczba liści nie miała dużego wpływu na wynik
- Minimalna liczba próbek na poziomie conajmniej 8 dawała lepsze wyniki

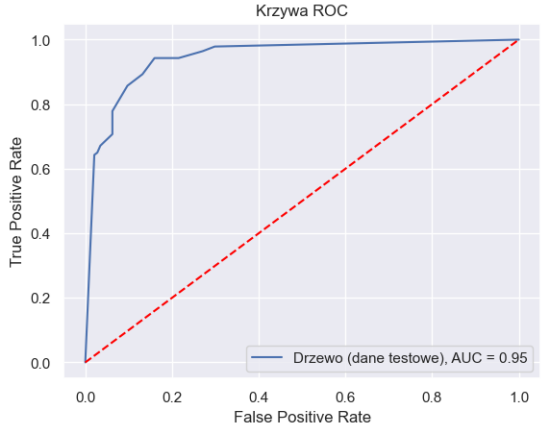
4 Najlepszy model

Wybierając najlepszy model postanowiłem zasugerować się wynikami przeprowadzonej kroswalidacji. Najlepszy model uzyskał AUC na poziomie 0.933, miał on następujące hiperparametry:

- $criterion = "entropy"$
- $max\_depth = 5$
- $min\_samples\_leaf = 8$
- $max\_leaf\_nodes = 20$

Sprawdźmy jak model radzi sobie w predykcji. Przyjrzyjmy się macierzy pomyłek, krzywej ROC oraz metrykom na zbiorach treningowym i testowym.



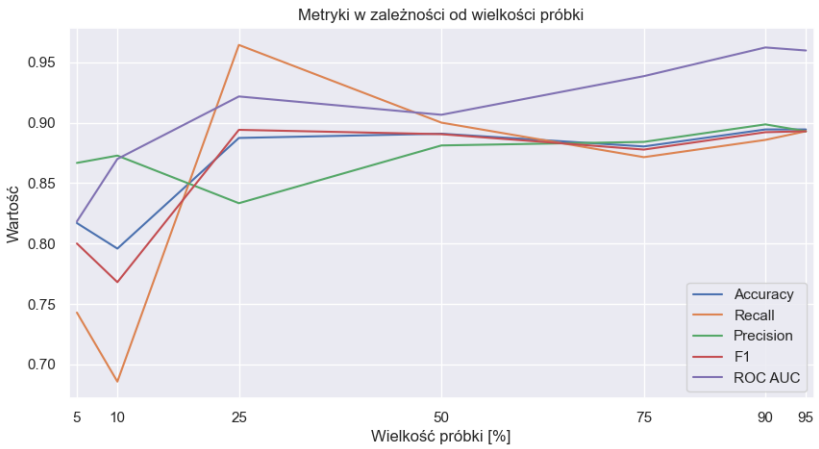


Metryka	Dane treningowe	Dane testowe
accuracy	0.93	0.89
recall	0.98	0.94
precision	0.89	0.85
f1-score	0.93	0.89
AUC	0.93	0.95

Widzimy, że model działa bardzo dobrze, zarówno na danych uczących, jak i testowych. Metryki są lepsze na zbiorze treningowym, natomiast nieznacznie, nie powinniśmy mieć obaw o przetrenowanie naszego modelu.

5 Wielkość próbki uczącej a skuteczność modelu

Ostatnim elementem zadania było sprawdzenie jak rozmiar zbioru treningowego wpływa na performens naszego drzewa. W tym celu wybierałem losowo próbki o zadanych rozmiarach z naszego pierwotnego zbioru treningowego, a następnie liczyłem metryki na całym zbiorze testowym. Uzyskałem takie wyniki:



Możemy zaobserwować, że rozmiar próbki pozytywnie wpływa na jakość predykcji, ale już próbka 25% pozwoliła osiągnąć zadowalające wyniki. Jednocześnie wyniki próbki 50% już praktycznie nie różnią się od tych dla całego zbioru. Oznacza to, że trenując nasz model moglibyśmy zmiejszyć ilość wykonywanych obliczeń poprzez zmiejszenie zbioru treningowego, a wciąż uzyskać bardzo dobry model drzewa.