Praca Domowa 1

Julia Girtler

2023-10-25

Cel Pracy

Przeprowadzenie eksperymetnu, który pokaże zależność dokładności od różnych parametrów funkcji DecisionTreeClassifier i wskaże najbardziej optymalne drzewo decyzyjne.

Przebieg eksperymentu

Została przperowadzona pięciokrotna kroswalidacja na zbiorze treningowych dla 160 drzew decyzjnych różniących się od siebie parmatrami takimi jak: criterion, max depth, min samples leaf, max leaf nodes. Powstała ramka danych która zawiera średnią dokładność dla każdego drzewa dla zbioru testowego oraz średnią dokłdność dla zbioru treningowego obliczona za pomocą kroswalidacji.

Przykładowe wiersze ramki danych

	Criterion	Max.Depth	Min.Samples.Leaf	Max.Leaf.Nodes	Train.Accuracy	Test.Accuracy
196	entropy	22	50	NA	0.8037143	0.8030000
2	gini	2	1	10	0.6672143	0.6726667
79	gini	15	100	20	0.7567857	0.7493333
157	entropy	10	100	5	0.6632857	0.6640000
25	gini	5	5	5	0.6685000	0.6720000

Obserwacje i wnioski

Criterion

Średnia dokładność dla "gini" jest większa niż dla "entropy". Wyniki te są bardzo zbliżone.

Criterion	Train.Accuracy	Test.Accuracy
entropy	0.7199521	0.7197283
gini	0.7218257	0.7268467

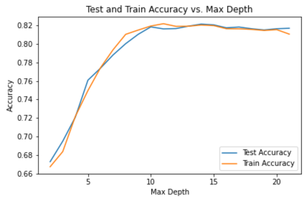
Max Depth

Ramka danych przedstawiająca średnią dokładność dla danej maksymalnej głębokości z uwzględnieniem innych parametrów.

Max.Depth	Train.Accuracy	Test.Accuracy
2	0.6671786	0.6705833
5	0.7205429	0.7257542
10	0.7383196	0.7393458
15	0.7393375	0.7404583
22	0.7390661	0.7402958

Możemy zauważyć, że im mniejsza maksymalna głebokość, tym dokładność na zbiorze testowym jest mniejsza, jendak przy wartosciach max depth powyżej 5 wyniki są bardzo zbliżone.Może być to spowodowane innymi parametrami, które również mogą wpływać na dokładność modelu.

Jeśli zbadamy drzewka gdzie będzie zmienny tylko parametr max depth, zaobserwujemy znaczący wzrost dokładności dla wartości max depth od 2 do 10.

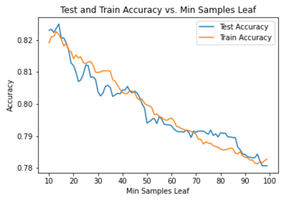


Min Samples Leaf

Ramka danych przedstawiająca średnią dokładność dla danej minimalnej ilości obserwacji w liściu z uwzględnieniem innych parametrów.

Min.Samples.Leaf	Train.Accuracy	Test.Accuracy
1	0.7230143	0.7250875
5	0.7230339	0.7256583
25	0.7213786	0.7241625
50	0.7199732	0.7217875
100	0.7170446	0.7197417

Jeśli zbadamy drzewka, gdzie będzie zmienny tylko parametr min samples leaf to możemy zaobserwować, że wraz ze wzrostem minimalnej ilości obserwacji w liściu spada dokładność.

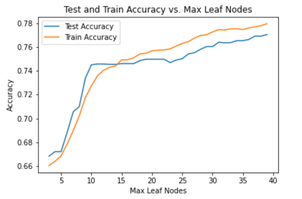


Max Leaf Nodes

Ramka danych przedstawiająca średnią dokładność dla danej maksymalnej ilości liści w drzewie z uwzględnieniem innych parametrów. Wraz ze wzrostem maksymalnej ilości liści wzrasta dokładność.

Max.Leaf.Nodes	Train.Accuracy	Test.Accuracy
5	0.6661500	0.6685167
10	0.7141643	0.7187167
20	0.7360543	0.7361033

Jeśli zbadamy drzewka gdzie będzie zmienny tylko parametr min samples leaf to możemy zaobserwować, że wraz ze wzrostem maksymalnej ilości liści wzrasta dokładność. Dynamika wzrotu jest większa dla mniejszych wartości max leaf nodes.



Wybór najlepszego drzewa

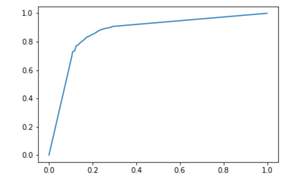
Największa dokładność na zbiorze testowym jest dla parametrów: criterion = "entropy", max depth = 15, min samples leaf = 5, max leaf nodes = None. Wynosi ona 0.8305. Przy kroswalidacji na zbiorze treningowym - 0.826786. Dla tego modelu przedstawione zostały macierz pomyłek i krzywa ROC.

Metric	Value
Accuracy	0.8190000
Recall	0.8134228
Precision	0.8205823
F1 Score	0.8169869

Macierz pomyłek

2502	518
505	2475

Krzywa ROC



Podsumowanie

Dobieranie odpowiednich parametrów funkcji DecisionTreeClassifier znacząco polepsza jakość drzewa decyzyjnego. Ich domyślne wielkości mogą nie być odpwiednie, ponieważ może dojść do przeuczenia modelu, co skutkuje obniżeniem dokładności dla testowego zbioru danych,