Praca domowa 1

Jan Kwiecień 320565

1 Wstęp

Celem pracy domowej numer 1 było sprawdzenie jak poszczególne parametry wpływają na jakość predykcyjną drzewa. Porównałem: minimalną liczbę obserwacji w liściu, głębokość drzewa, kryterium podziału oraz liczbę losowo wybieranych cech potrzebnych do decyzji co do podziału (parametr *max_features*).

2 Eksperyment

2.1 Dobór parametrów

W pierwszej kolejności podzieliłem zbiór trenignowy i testowy w proporcji 7:3. Opisane powyżej parametry dobrałem w następujących skalach:

- kryterium podziału (*criterion*): *gini*, *entropy*;
- minimalna liczba obserawacji w liściu (*min_samples_leaf*): 50, 100, 150;
- głębokość drzewa (*max_depth*): 1-10;
- liczba losowo wybieranych cech potrzebnych do podziału (*max_features*): 1 - 19.

2.2 Sposób testowania

W pierwszej części użyłem dwóch pętli (w jednej kryterium podziału *gini*, w drugiej *entropy*) do znalezienia najbardziej optymalnych parametrów *min_samples_leaf* oraz *max_depth*. Następnie po odpowiednim doborze napisałem kolejne dwie pętle (z takim samym podziałem co do kryteriów), aby przetestować parametr *max_features*. W każdym z czterech przypadków używałem pięciokrotnej kroswalidacji. Miarą dokładności w opisanym teście było *roc_auc* (pole pod krzywą *roc*).

2.3 Wyniki

Po pierwszym teście zarówno w kryterium podziału *gini* jak i *entropy* najlepsze okazały się być parametry: *max_depth* = 10, *min_samples_leaf* = 50. *Roc_auc* dla każdego z nich prezentowało się następująco:

- *gini*: 0.8740553816488278
- *entropy*: 0.8810291529926146

Następnie w drugim teście najbardziej optymalny parametr *max_features* dla kryterium podziału *gini* wyniósł 19, a dla *entropy* 16. Ostateczne *roc_auc* dla każdego z nich:

- *gini*: 0.8755300333312832
- *entropy*: 0.8824244567096807

2.4 Wybór drzewa

Biorąc pod uwagę największą wartość *roc_auc*, ostatecznie do kolejnego zadania wybrałem drzewo decyzyjne z następującymi parametrami:

- *criterion*: *entropy*
- *min_samples_leaf*: 50
- *max_depth*: 10
- *max_features*: 16

3 Analiza jakości predykcyjnej modelu

W celu dokonania analizy jakości predykcyjnej modelu wyznaczyłem: macierz pomyłek, dokładność, czułość i precyzję. Ponadto narysowałem krzywą ROC i policzyłem wartość AUC.

3.1 Macierz pomyłek

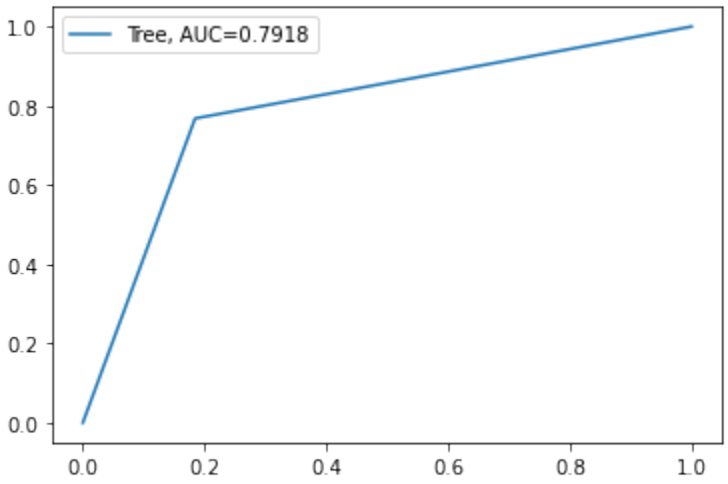
Macierz pomyłek prezentuje się następująco: $CF = \begin{bmatrix} 2450 & 554 \\ 695 & 2301 \end{bmatrix}$

3.2 Miary dla otrzymanych predykcji

- dokładność: 0.7918333333333333
- czułość: 0.7680240320427236
- precyzja: 0.805954465849387

3.3 Krzywa ROC i AUC

Krzywa ROC:



Wartość AUC: 0.79

4 Wnioski

Wybrane przeze mnie drzewo okazło się być skuteczne na zbiorze testowym. Otrzymałem wysoką dokładność, czułość, precyzję i wartość AUC. Macierz pomyłek również wskazuje na wysoką efektywność modelu.