Sprawozdanie PD1 WdUM

Jakub Niemyjski

24 października 2023

1 Cel projektu

Głownym celem tego projektu jest zbadanie wpływu zmiany parametrów na jakość predykcyjną modelu.

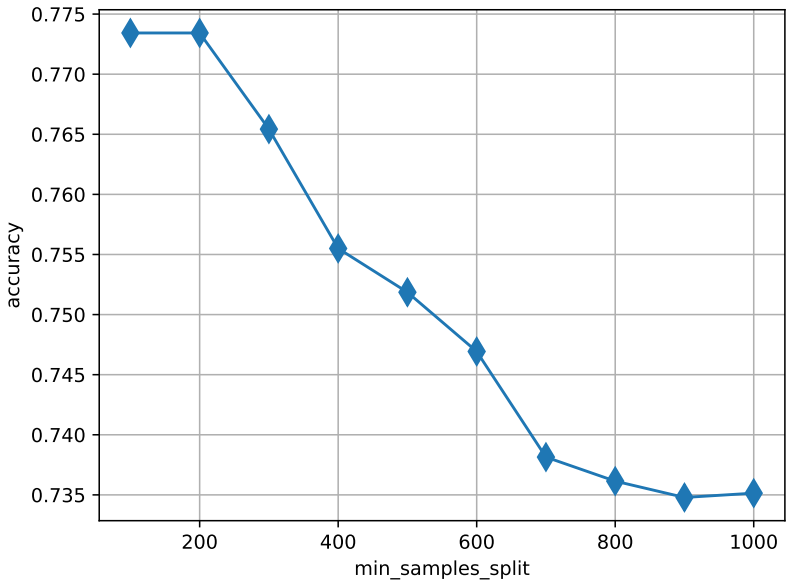
2 Eksperyment

W przygotowanym eksperymencie ukazującym miarę dokładności drzewa w zależności od trzech parametrów drzewa parametry przyjmowały następujące wartości:

- kryterium podziału *(criterion)*: `'gini', 'entropy'`
- głębokość drzewa *(max_depth)*: $\mathbb{N} \cap [10, 15)$
- minimalna liczba obserwacji w liściu *(min_samples_leaf)*: $\{50, 60, \ldots, 100\}$.

Ostatecznie, po dokonaniu pięciokrotnej kroswalidacji dla każdej kombinacji powyższych trzech parametrów `DecisionTreeClassifier` wybrany został model z parametrami, dla których średnio (w sensie średniej arytmetycznej) model ten cechował się najlepszą dokładnością.

Przy tak ustalonych parametrach, zbadano jeszcze wpływ czwartego parametru (*min_samples_split* o wartościach ze zbioru $\{10, 200, \ldots, 1000\}$) na dokładność otrzymanego modelu, ponownie za pomocą kroswalidacji. W związku z przedstawionym wykresem poniżej widać, że skuteczność jest najwyższa dla *min_samples_split* $= 100$ lub 200.



Rysunek 1: Zależność dokładności modelu w zależności od parametru *min_samples_split*

Następnie utworzono model o trzech pierwszych parametrach dostosowanych jak wcześniej i parametrem *min_samples_split* równym 100, po czym nauczono go na zbiorach `X_train` oraz `y_train`.

Uzasadnienie wyboru parametrów

Oczywiście, najefektywniejszy model możnaby tworzyć na inne sposoby w zależności od tych badanych czterech parametrów, jednak, przykładowo, algorytm badający dokładności modelów dla każdej możliwej kombinacji tych parametrów byłby znacznie bardziej złożony czasowo, dlatego ograniczono się tutaj do badania każdej kombinacji pierwszych trzech parametrów, a czwarty był już dostosowany pod ustalone pierwsze trzy. Podobnie, możnaby zwiększyć liczbę możliwych przeszukiwanych wartości badanych parametrów, jednakże ponownie zachodzi tu znaczne zwiększenie długości wykonywania takiego algorytmu.

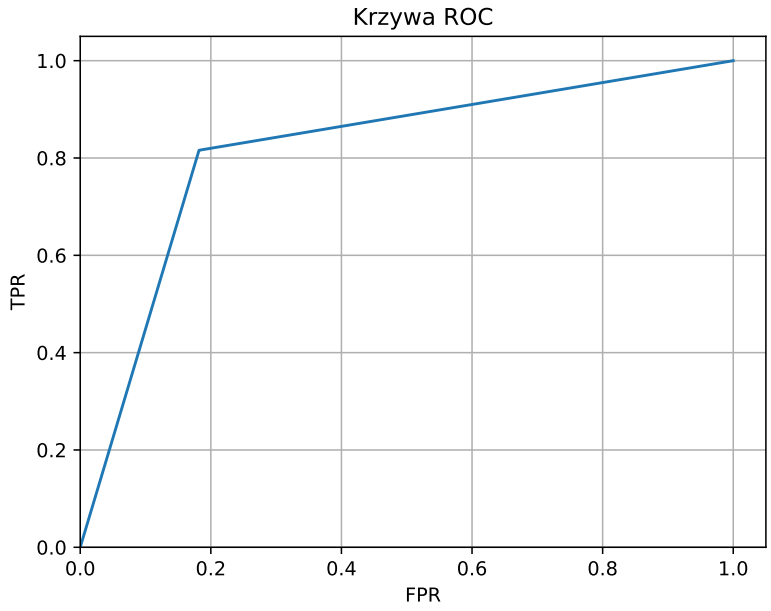
3 Analiza jakości predykcyjnej modelu

Dla wybranego w poprzednim punkcie modelu macierz pomyłek na zbiorze testowym jest następująca:

$$\begin{bmatrix} tn & fp \\ fn & tp \end{bmatrix} = \begin{bmatrix} 2473 & 550 \\ 548 & 2429 \end{bmatrix},$$

gdzie t - true, f - false, p - positive, n - negative. Dla takiej macierzy dokładność to 0.8170, czułość to około 0.8159, precyzja to około 0.8154.

Wyznaczono również krzywą ROC dla tego modelu, dla której AUC wynosi około 0.8170. Krzywą tą przedstawiono poniżej.



4 Podsumowanie

Patrząc na wykres 1 może się wydawać, że spadek dokładności wraz ze wzrostem parametru *min_samples_split* jest nieznaczny. Na tym wykresie dokładność waha się między, około 0.735, a 0.773. Jednak, przykładowo, przy przeprowadzeniu klasyfikacji dla 1000 rekordów oznacza to, że średnio o $773 - 735 = 38$ więcej przypadków zostanie w gorszym z nich sklasyfikowane niepoprawnie.

Istotną kwestią jest również to, czy przewidziana dokładność 0.8170 jest dla nas satysfakcjonująca. Jest to jednak pytanie, na które nie da się udzielić jednoznacznej odpowiedzi z powodu braku kontekstu

otrzymanych danych. Można wyobrazić sobie sytuację, gdy warunkiem koniecznym akceptacji takiego modelu byłoby posiadanie dokładności wyznaczonej na zbiorze testowym większej niż, przykładowo, 0.90, a tego już nasz model nie spełnia, co oznaczałoby, że musimy z niego zrezygnować. Przykładem takiej sytuacji może być wykrywanie groźnej choroby, przeciw której podjęcie leczenia wiązałoby się również z negatywnymi skutkami dla zdrowia, jednak decyzja o tym, jaki poziom dokładności modelu jest cennym modelem powinnno być zostawione osobom kompetentym w dziedzinie, której ten model dotyczy.