Raport

Tymoteusz Urban

26 czerwca 2023

1 Opis rozwiązania

Moje rozwiązanie jest krótkie i proste, jednak zawiera wszystkie ważne elementy konstrukcji modelu predykcyjnego. Na początku zająłem się eksploracją danych - zapoznałem się z kolumnami, typami i brakami danych. Używając histogramów i funkcji value_counts podjąłem decyzje, które kategoryczne kolumny wymagają zakodowania, a które należałoby usunąć. Po zastosowaniu one hot encodingu zbadałem korelację pomiędzy zmiennymi i usunąłem te, które były mocno skorelowane i nie wydawały się kluczowe dla celu biznesowego.

Ponadto w preprocessingu zająłem się także brakami danych, które zastąpiłem medianami. Napisałem także funkcję usuwającą outliery, jednak zdecydowałem się z niej nie korzystać, gdyż outliery w obszarze cyberbezpieczeństwa mogą być kluczowe w ocenie zagrożenia. Dane poddałem także standaryzaji, jednak ostateczne wyniki są bez jej użycia, gdyż zwyczajnie były lepsze bez używania żadnego skalera.

Wreszcie przeszedłem do etapu tworzenia modelu. Sprawdziłem trzy klasyfikatory, które zwykle dają jedne z najlepszych wyników: regresja logistyczna, xg boost i lasy losowe. Najlepsze wyniki uzyskał xgboost, zatem zdecydowałem się go poddać procesowi strojenia hiperparametrów. Użyłem do tego funkcji randomizedserach z walidacją krzyżową (optymalizowałem wartość auc score). Ostateczny model uzyskał wartość auc na poziomie 0.902 na ziorze testowym (wyodrębnionym z pierwotnego treningowego).

2 Kod rozwiązania

Link do repozytorium na githubie: qedsoftware_recruation