Airline passenger satisfaction

06-DUMALIO 2022/Z

Cel projektu

Celem projektu było stworzenie modelu, który przewiduje, czy pasażer jest zadowolony z linii lotniczej na podstawie płci, typu klienta, wieku, typu podróży, klasy podróży, odległości lotu, usługi Wi-Fi podczas lotu, dogodność godziny wylotu/przylotu, łatwości w rezerwacji biletu online, lokalizacji bramki, jakości jedzenia i picia, możliwość boardingu online, komfortu siedzenia, zapewnianych rozrywek podczas lotu, jakości obsługi pokładowej, jakości usługi check-in, obsłudze bagażu, czystości, opóźnienia odlotu, opóźnienia przylotu.

Dane

Dane pochodzą z platformy Kaggle.com
link - https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Po połączeniu train.csv i test.csv zbiór składa się z 129880 przykładów. Dane zostały podzielone w stosunku 0.8 x 129880 = 103904 przykładów w zbiorze treningowym i 25976 przykładów w zbiorze testowym.

Modele

Dane wejściowe zostały znormalizowane przy wykorzystaniu narzędzie StandardScaler z biblioteki scikit-learn.

W projekcie porównano działanie 5 modeli:

- Dwuklasowa regresja logistyczna.
- Algorytm k najbliższych sąsiadów w którym najpierw ustawiony został parametr n_neighbors na liczbę 3, a następnie z wykorzystaniem GridSearchCV znaleziona została optymalna wartość tego parametru równa 5. Więc ostatecznie wykorzystaliśmy algorytm k najbliższych sąsiadów z n_neighbors równym 5.
- Klasyfikator drzewa decyzyjnego, którego parametr criterion pozostał na wartości domyślnej „gini” oraz parametr splitter pozostał na swojej domyślnej wartości „best”.
- Naiwny klasyfikator bayesowski. Jako modelu klas użyto rozkładu normalnego.
- Sieć neuronowa wykorzystująca model Sequential z biblioteki Keras w której algorytmem optymalizacji jest optimizer ‘adam’, funkcją straty jest funkcja ‘binary_crossentropy’ a metryką, która używana jest do oceny jakości prognoz sieci jest metryka ‘accuracy’

Ewaluacja

Do ewaluacji wykorzystano metryki *accuracy, precision, recall* i *F1-score*. Wyniki ewaluacji przedstawia poniższa tabelka:

Model	Accuracy	Precision	Recall	F1-score
Dwuklasowa regresja logistyczna	0.87384508777332	0.87344155670507	0.86951804943202	0.87118182634916
Algorytm k najbliższych sąsiadów	0.9278564829072	0.9313709436613	0.92266090749044	0.92602431090476
Naiwny klasyfikator bayesowski	0.8662996612257	0.8663527545108	0.86116231433630	0.86325663908432
Klasyfikator drzewa decyzyjnego	0.9424853711117	0.9414787680866	0.94166866825458	0.94157301164925
Sieć neuronowa	0.9535725284878	0.9552780594663	0.95055229731817	0.95259507390715

Wnioski

Najlepsze wyniki pod względem F1-score uzyskano przy pomocy sieci neuronowej. Najgorzej pod tym względem wypada naiwny klasyfikator bayesowski, którego F1-score wynosi około 0.86, blisko tego wyniku jest również dwuklasowa regresja logistyczna, której F1-score wynosi około 0.87. Warto zauważyć, że zarówno accuracy, precision, recall oraz F1-score w sieci neuronowej w żadnej z tych metryk nie mamy wyniku poniżej 0.95.