Projekt 1 - raport

Mateusz Grzyb, Bartosz Eljasiak

Poniższy raport zawiera informacje na temat tego, co zostało przez nas zrobione, podczas trzech etapów projektu, oraz jakie są tego wyniki.

Ponieważ wszystkie wyniki dostępne są w plikach .ipynb, w tym dokumencie zawrzemy najbardziej istotne, naszym zdaniem, wnioski.

Zbiór danych

Nazwa: sick

Autor: Ross Quinlan

Opis:

"Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Syndney, Australia. 1987."

Źródła:

- https://datahub.io/machine-learning/sick#data
- http://archive.ics.uci.edu/ml/datasets/thyroid+disease

Etap 1

Na pierwszy etap projektu składa się przede wszystkim eksploracyjna analiza danych.

Co zostało zrobione:

- zbadane zostały typy zmiennych i braki danych
- przeprowadzone zostało wstępne czyszczenie danych
  - rzutowanie typów str na bool
  - usunięcie redundantnych kolumn (informujących o obecności pomiaru)
  - usunięcie kolumny bez danych (TBG)
  - zastąpienie błędnej obserwacji wieku niewiadomą (w kolumnie age)
- przeprowadzona została analiza zmiennych
  - rozkłady zmiennych kategorycznych
  - rozkłady zmiennych liczbowych
  - statystyki pozycyjne i rozproszenia zmiennych liczbowych
- oszacowane zostały odsetki badanych, u których poziom TSH przekracza dolną, bądź górną normę
- zbadane zostały korelacje zmiennych
- przeprowadzona została analiza wielowymiarowa
  - "Jaki jest odsetek chorych, w danym przedziale wiekowym (z podziałem na płeć)?"
  - "Jakie są korelacje poziomów hormonów oraz diagnozy?"
  - "Jakie są rozkłady poziomów hormonów wewnątrz danej diagnozy?"

Najważniejsze wnioski:

- dane obejmują 3772 pacjentów
- zmienne są typów int, float i str (większość z pośród ostatnich może być rzutowana na bool)
- występują braki danych (dotyczące płci i poziomów hormonów)
- około 2/3 badanych to kobiety
- chorzy stanowią zdecydowaną mniejszość badanych (zbiór jest niezbalansowany)
- rozkłady zmiennych liczbowych (prócz zmiennej TSH) są "ładne"
- dla zmiennej TSH korzystne może być wykonanie przekształcenia logarytmicznego
- niektóre pary poziomów hormonów wykazują silną korelację
- w niemal każdym przedziale wiekowym odsetek chorych jest większy dla mężczyzn
- odsetek chorych rośnie wraz z wiekiem
- hormon T3, poniżej pewnego poziomu, zwiastuje chorobę tarczycy z dużym prawdopodobieństwem
- rozkłady poziomów hormonów różnią się w zależności od diagnozy
- rekordowe wartości TSH dotyczą osób zdrowych

Etap 2

Na drugi etap projektu składa się inżynieria cech i wstępne uczenie maszynowe.

Co zostało zrobione:

- przeprowadzone zostało wstępne czyszczenie danych (podobnie jak w etapie 1)
- przeprowadzona została dyskusja nad wątpliwą jakością zmiennej referral_source
- przeprowadzona została imputacja zmiennej sex (w sposób losowy z prawdopodobieństwami odpowiadającymi rozkładowi płci w danych)
- z pomocą krótkich testów wybrane zostały algorytmy do sprawdzenia (RandomForest, XGBoost), a także metody walidacji (StratifiedKFold) i metryki (ROC curve, AUC, F1)
- napisane zostały specjalne funkcje testujące, wykorzystujące kroswalidację i wspomniane wcześniej miary
  - funkcja cv_roc, kreśląca krzywą ROC dla każdej iteracji, a także formie uśrednionej, z przedziałem niepewności, oraz obliczająca AUC dla każdej iteracji, a także w formie uśrednionej, z odchyleniem standardowym
  - funkcja cv_f1, obliczająca wynik F1 dla każdej iteracji, a także w formie uśrednionej, z odchyleniem standardowym, oraz przedstawiająca wyniki na pionowym wykresie pudełkowym
- z pomocą wyżej opisanych funkcji przetestowane zostały wybrane algorytmy, dla różnych sposobów imputacji (usunięcie wybrakowanych kolumn, imputacja średniej, imputacja mediany, Iterative Imputer z pakietu sklearn) (w celu uniknięcia przecieku danych wykorzystano pipe'y z pakietu sklearn)
- opisane powyżej testy przeprowadzone zostały ponownie, po uprzednim usunięciu wątpliwej zmiennej referral_source
- opisane powyżej testy przeprowadzone zostały ponownie, po uprzedniej dalszej redukcji danych (kolumny do usunięcia wybrane zostały z pomocą testu korelcji chi-kwadrat) (dane zredukowano z 30 do 12 kolumn)

Najważniejsze wnioski:

- wyniki kroswalidacji wykazały poprawność metodyki
- najlepsze wyniki uzyskał XGBoost z imputacją mediany
- usunięcie zmiennej referral_source, pomimo jej wysokiej korelacji ze zmienną celu, marginalnie pogarsza wyniki modeli
- usuwanie wybrakowanych kolumn powoduje niemal całkowity brak predykcji typu TP (miara AUC jest tu nieco złudna, ale dobrze widać to po wyniku F1)
- wymiarowość danych można znacznie zredukować i wciąż uzyskiwać bardzo dobre wyniki
- XGBoost jest algorytmem szybszym, niż algorytm Random Forest

Etap 2

Na drugi etap projektu składa się strojenie hiperparametrów i Feature Importance.

Co zostało zrobione:

- przeprowadzone zostało wstępne czyszczenie danych (podobnie jak w etapie 2, jednak tym razem pozostawiono kolumny "*_measured" i usunięto zmienną referral_source)
- dane podzielone zostały na zbiór treningowy (do strojenie hiperparametrów) i testowy (do testowania najlepszego modelu na wcześniej nieużywanych danych)
- przygotowana została specjalna metryka Weighted TPR-TNR, bazująca na nowym artykule naukowym (https://www.sciencedirect.com/science/article/abs/pii/S0957417420302153?via%3Dihub) (w celu użycia jej przy strojeniu hiperparametrów)
- przeprowadzone zostało strojenie hiperparametrów
  - użyto funkcji RandomizedSearchCV
  - zakresy parametrów dobrano starannie na bazie kilku artykułów oraz dokumentacji
  - wykorzystano wcześniej wspomnianą metrykę Weighted TPR-TNR
- najlepszy model porównanany został do domyślnego, na podstawie Weighted TPR-TNR oraz MCC (Matthews correlation coefficient)
- przygotowana została macierz pomyłek dla najlepszego modelu
- przygotowana została analiza Feature Importance dla najlepszego modelu (Weight, Gain i Coverage)

Najważniejsze wnioski:

- metryka Weighted TPR-TNR rzeczywiście bardzo dobrze sprawdza się dla zbiórów niezbalansowanych (ocena subiektywna, wynikająca z zestawienia wartości metryki z macierzą pomyłek)
- metryka MCC również stanowi dobrą alternatywę, ponadto jest już dostępna w pakiecie sklearn
- strojenie hiperparametrów pozwoliło zauważalnie poprawić wyniki modelu
- wyniki najlepszego modelu na zbiorze testowym są bardzo dobre
  - Weigted TPR-TNR 0.913 (dla domyślnego 0.856)
  - MCC 0.845 (dla domyślnego 0.819)
  - model tylko w 1.7% przypadków fałszywie poinformował o chorobie i jednocześnie wykrył chorobę w 93% przypadków
- zmienna FTI występuje w drzewach najczęściej (a miały one głębokość ≤ 6), a zmienna T3 dała najwyższy gain (znacznie wyróżnia się na tle pozostałych)