Projekt 2

Twoim zadaniem jest zbudowanie modelu predykcyjnego, w oparciu o rzeczywisty zbiór danych – listę pasażerów „Titanica". Model ten, w oparciu o zmienne, wybrane spośród 12 predyktorów (atrybutów predykcyjnych) pozwoli nam wyznaczyć wartość zmiennej celu. Atrybutem, którego wartość będziemy przewidywać, jest „survival", który mówi nam czy dany pasażer przeżył katastrofę, czy też nie. Sam zbiór danych dostępny jest w pakiecie *scikit-learn*, możesz go też należć za pomocą Google (plik „titanic.csv")[1] i załadować do swojego projektu. Zawiera on 1309 rekordów, opisywanych przez następujące atrybuty[2]:

- *survival* – informacja, czy pasażer przeżył (1 – tak, 0 – nie)
- *class* – klasa kajuty, którą pasażer podróżował (1 – najwyższa, 3 – najniższa)
- *name* – imię, nazwisko oraz tytuł pasażera
- *sex – płeć pasażera*
- *age* – wiek pasażera
- *sibsp* – liczba rodzeństwa oraz małżonków, podróżujących wraz z pasażerem
- *parch* – liczba dzieci oraz rodziców, podróżujących wraz z pasażerem
- *ticket* – numer biletu pasażera
- *fare* – cena biletu pasażera
- *cabin* – numer kajuty, w której pasażer podróżował
- *embarked* – port zaokrętowania pasażera (C – Cherbourg, Q – Queenstown, S – Southampton)
- *boat* – numer łodzi ratunkowej, w której pasażer się ewakuował
- *body* – numer identyfikacyjny znalezionego ciała.

Na początku powinieneś przyjąć pewne założenia odnośnie modelu, np. kiedy należy zakończyć „zabawę" nad jego udoskonalaniem. Jako punkt odniesienia przyjmijmy zwykle osiągany w tego typu analizach poziom trafności (ang. *accuracy*) rzędu 80%. Wynik nie gorszy, niż ta liczba uznaj za sukces i zakończ pracę nad projektem.

1. Analiza zbioru danych

Zacznij od pobrania danych, a następnie załaduj plik „titanic.csv" do przestrzeni roboczej swojego projektu. By przyjrzeć się nieco bliżej danym, dokonaj sumaryzacji danych. Pozwoli Ci to podejrzeć podstawowe metryki dla każdej kolumny, np. wartości minimalne, wartości maksymalne, średnią, medianę, itd.

Spróbuj odnaleźć błędy w danych – brakujące wartości, błędna interpretacja rodzaju zmiennych itp. Zastanów się, które atrybuty wybrać do analizy (pod kątem ich istotności dla przewidywań modelu, jakości, czy też tzw. wycieków danych), czy i jak uzupełnić brakujące dane, dokonać ich transformacji itd.

[1] Pełen zbiór danych dostępny jest np. tu: https://analityk.edu.pl/wp-content/uploads/2020/02/titanic.csv
[2] W oczywisty sposób użycie informacji z atrybutów „boat" i „body" spowodowałoby tzw. wyciek danych, co pozwoliłoby modelowi nauczyć się tego, czego w przeciwnym razie by „nie wiedział".

2. Feature Engineering

Na przykład, w kwestii uzupełniania wartości kolumny „age", po analizie danych możesz dojść do kilku wniosków, które przedstawiają się następująco:

- Wszystkie kobiety o tytule „Miss", które nie mają dzieci, to damy o średnim wieku … lat.
- Panowie noszący tytuł „Master", to kawalerzy, których średnia wieku to ok. … roku.
- „Sir", „Mr", „Ms" i „Mrs" to z małymi wyjątkami dojrzałe osoby. Wiek dla wszystkich z nich uzupełnij, posiłkując się średnią dla danej grupy.
- Wszyscy panowie z tytułem „Dr", to dojrzali mężczyźni o średniej wieku … lata.

Wiek dla wszystkich powyższych grup uzupełnij, posiłkując się średnią dla danej grupy. Podobnie, dla wszystkich innych osób, które nie mają podanego wieku, uzupełnij ich wiek średnią dla danej płci.

Utwórz trzy dodatkowe zmienne, których użyjesz w dalszej części eksperymentu:

„family_size" – poprzez dodanie do siebie dwóch istniejących wartości: „parch" i „sibsp", oraz liczby 1, odpowiadającej za daną osobę. Intuicyjnie można przyjąć, że wielkość rodziny mogła mieć znaczący wpływ na to czy komuś udało się przeżyć, czy też nie.

„age_range" – jest to zmienna kategoryczna, która przypisuje pasażera do jednej z czterech kategorii wiekowych: „Bobas", „Dzieciak", „Nastolatek", „Dorosly". Mamy tu na uwadze zasadę „women and children first". Przyjmij wiek 6, 12 i 18 lat jako punkty podziału zbioru.

„mpc" – zmienna, która ma „uwypuklić" szanse na przeżycie dzieci i osób z pierwszej klasy. Jest ona wynikiem mnożenia wieku danej osoby i klasy, w której podróżowała. Dla przykładu, 5-letnie dziecko podróżujące w pierwszej klasie (wynik = 5), miało dużo większe szanse na przetrwanie katastrofy niż 70 letni pan podróżujący klasą trzecią (wynik = 210).

Te trzy dodatkowe zmienne będą dobrym punktem wyjścia do walki o jak najwyższy wynik Twojego modelu.

3. Edycja metadanych

Przed przejściem do uzupełnienia brakujących danych należy sprawdzić (i ew. zmienić) typy poszczególnych kolumn. Zwróć w szczególności uwagę na kolumny: „survived", „pclass", „embarked" i „sex" – powinny one reprezentować dane kategoryczne (ang. *factors*). Jakiego typu wartości powinna zawierać kolumna „fare"?

4. Uzupełnienie brakujących danych

Pozostaje Ci teraz uzupełnić brakujące wartości dla pozostałych atrybutów. Można posłużyć się tu jedną z metod tzw. imputacji, zawartą w pakiecie *scikit-learn* (https://scikit-learn.org/stable/auto_examples/impute/plot_missing_values.html) – w szczególności dobre efekty daje metoda MICE (ang. *Multivariate Imputation by Chained Equations*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/), implementowana poprzez klasę *sklearn.impute.IterativeImputer* (https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html).

5. Przycięcie odstających wartości

Przycięcie odstających wartości zacznij od analizy wartości poszczególnych atrybutów. Najlepszym do tego narzędziem jest wykres punktowy, gdzie nanosimy na obu osiach ten sam atrybut. Dzięki temu czarno na białym widać, które wartości odstają. Na przykładzie wizualizacji wieku poszczególnych pasażerów, możemy zauważyć, że dla kolumny „age" powinniśmy usunąć wszystkie wartości powyżej 67 i zastąpić je średnim wiekiem. Podobnie należy postąpić z ceną biletu, jaką zapłacili poszczególni pasażerowie. Odstające wartości zostają zastąpione średnią. UWAGA: Przycinania odstających wartości zawsze dokonujemy przed normalizacją danych.

6. Normalizacja danych numerycznych

Wszystkie dane numerycznie muszą zostać znormalizowane. Dlaczego? Otóż podczas uczenia maszynowego zmienne numeryczne o większych wartościach mogą być postrzegane przez algorytm jako ważniejsze. By lepiej to wyjaśnić, posłużmy się tu przykładem z danych dotyczących pasażerów Titanica. Dana pasażerka o imieniu „Sandstrom, Miss. Marguerite Rut" szczęśliwie przeżyła katastrofę „Titanica". Z danych jasno wynika, że zapłaciła za bilet 16.7, a w chwili podróży miała jedynie 4 lata. Algorytm uzna, że zdecydowanie większy wpływ na to, że udało jej się ocaleć miała cena biletu, co oczywiście nie musi być prawdą. Chyba każdy (zwłaszcza po obejrzeniu filmu „Titanic" Jamesa Camerona) jest w stanie postawić hipotezę, że największe szanse na przeżycie miały kobiety i dzieci… Właśnie dlatego należy dokonać normalizacji.

Oczywiście, normalizacji powinniśmy dokonać dla zmiennych liczbowych (nie chcemy, by algorytm dokonywał normalizacji danych kategorycznych). Jako metodę użytą do transformacji danych wybierz „MinMaxScaler" lub „StandardScaler".

7. Wybór algorytmu

Bez wątpienia mamy tu do czynienia z dwuklasowym problemem klasyfikacyjnym. W oparciu o podane atrybuty musimy wyznaczyć jedną z dwóch wartości: [0, 1], [Tak, Nie], [Prawda, Fałsz], itd. W naszym wypadku zmienną celu, której wartość chcemy przewidzieć, będzie kolumna „survived".

Finalnie powinieneś wybrać maksymalnie 6-8 atrybutów które posłużą Ci jako predyktory. Z doświadczenia wiadomo, że przy takiej liczbie atrybutów, przy problemie klasyfikacyjnym najlepsze rezultaty daje drzewo decyzyjne lub las drzew losowych. Nie kończ jednak na tym poszukiwania „idealnego" algorytmu – spróbuj innych, dostępnych w pakiecie *scikit-learn* metod klasyfikacji.

By porównywać wyniki jakie dają poszczególne algorytmy musisz podzielić dane na uczące i walidujące. Dane podziel w sposób losowy, przy czym niech 80% stanowią dane uczące. Następną kwestią jest wybór zmiennych predykcyjnych – spróbuj na początek użyć kolumn „sex", „age", „age_range", „pclass", oraz „fare". Pozostaje Ci teraz nauczenie wybranych modeli (algorytmów) na danych treningowych (wykorzystaj podzbiór walidacyjny do tego, by sprawdzić, czy nie nastąpiło przeuczenie modelu), zweryfikowanie ich działania na danych testowych oraz porównanie otrzymanych wyników.