# Metody klasyfikacji

Jakub Zbrzezny

Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

09.06.2020

# Plan prezentacji

- Cel projektu
- Przygotowanie danych
- Sprawdzenie metod klasyfikacji
- Podsumowanie

# Wprowadzenie

Celem projektu jest przewidzenie, czy dane gospodarstwo dotyka bieda. Gospodarstwo jest objęte biedą, jeżeli wartość zmiennej *poor* wynosi "Poor".

Zmienna *poor* występuje w zbiorach o nazwach kończących się na *labels* - jest to więc zmienna objaśniana.

Natomiast zmiennymi objaśniającymi są wszystkie zmienne z odpowiadającego zbioru o nazwie kończącej się na *data* - jest ich bardzo dużo, więcej niż 100.

# Przygotowanie danych

Złączyłem zbiór *train data*, *train labels* w jeden zbiór oraz **test data**, *valid labels* również w jeden zbiór.

Następnie łączyłem oba otrzymane zbiory, żeby czynniki były takie same, a następnie indeksowałem na zbiór treningowy oraz testowy.

Dodatkowo jeszcze sprawdziłem, gdy łączyłem zbiór treningowy z walidacyjnym.

# Sprawdzenie metod klasyfikacji - 1 klasyfikator

1 klasyfikatorem jest *randomForest* z liczbą drzew równą 50.

Robiłem model pełny (zmienną poor uzależniłem od reszty). Prawdopodobieństwo, że predykcja zmiennej *poor* równa jest zmiennej *poor* ze zbioru testowego: 0.5006667 dla zbioru testowego oraz 0.7766667 dla zbioru walidacyjnego.

AUC dla zbioru testowego: 0.499
AUC dla zbioru walidacyjnego: 0.862

# Sprawdzenie metod klasyfikacji - 2 klasyfikator

2 klasyfikator jest *randomForest* z liczbą drzew równą 50 wraz z selekcją zmiennych funkcją *CMIM* z parametrem $k = 6$.
Dla testowego zostały wybrane następujące zmienne:
*cons 0901*, *cons 0106*, *geo district*, *cons 0501*, *cons 0111*, *cons 0508*
Dla walidacyjnego wybrane zmienne były inne niż dla testowego.

Prawdopodobieństwo dla zbioru testowego: 0.5056667
Prawdopodobieństwo dla zbioru walidacyjnego: 0.7493333

AUC dla zbioru testowego: 0.504
AUC dla zbioru walidacyjnego: 0.820

# Sprawdzenie metod klasyfikacji - 3 klasyfikator

3 klasyfikator jest ten sam, co wcześniej, ale z liczbą zmiennych równą 12.
Dla testowego zostały wybrane następujące zmienne:
*cons 0901*, *cons 0106*, *geo district*, *cons 0501*, *cons 0111*, *cons 0508*, *hld nbcellpho*, *cons 0401*, *cons 0801*, *cons 1108*, *cons 1204*, *hld headsleep*
Dla walidacyjnego wybrane zmienne były inne niż dla testowego.

Prawdopodobieństwo dla zbioru testowego: 0.5
Prawdopodobieństwo dla zbioru walidacyjnego: 0.7523333

AUC dla zbioru testowego: 0.500
AUC dla zbioru walidacyjnego: 0.840

# Sprawdzenie metod klasyfikacji - 4 klasyfikator

Teraz sprawdzę działanie *randomForest* wraz selekcją, korzystając z danych indywidualnych.
Wziąłem pierwsze 7 kolumn z danych indywidualnych.
W 6, 7 kolumnie były braki danych, więc wiersze z brakami danych usunąłem.
Zmienne dla testowego (dla walidacyjnego były inne):
*cons 0801*, *geo district*, *cons 0111*, *cons 0106*, *hld nbcellpho*, *cons 0508*

Prawdopodobieństwo dla zbioru testowego: 0.6616667
Prawdopodobieństwo dla zbioru walidacyjnego: 0.754

AUC dla zbioru testowego: 0.494
AUC dla zbioru walidacyjnego: 0.495

# Sprawdzenie metod klasyfikacji - 5 klasyfikator

5 klasyfikatorem jest *bagging* z parametrem *mfinal* $= 25$, z modelem pełnym.

Prawdopodobieństwo dla zbioru testowego: 0.4973333
Prawdopodobieństwo dla zbioru walidacyjnego: 0.7543333

AUC dla zbioru testowego: 0.495
AUC dla zbioru walidacyjnego: 0.835

# Sprawdzenie metod klasyfikacji - 6 klasyfikator

6 klasyfikatorem jest *bagging* z selekcją funkcją *JMIM* z parametrem $k = 12$.
Dla zbioru testowego otrzymałem następujące zmienne (dla walidacyjnego znów były inne): *cons 0901*, *hld dwateros*, *geo district*, *cons 0106*, *cons 0111*, *cons 0501*, *cons 0801*, *hld nbcellpho*, *cons 0401*, *hld cooking*, *cons 0803*, *cons 1204*

Prawdopodobieństwo dla zbioru testowego: 0.498
Prawdopodobieństwo dla zbioru walidacyjnego: 0.7403333

AUC dla zbioru testowego: 0.492
AUC dla zbioru walidacyjnego: 0,82

# Sprawdzenie metod klasyfikacji - 7 klasyfikator

7 klasyfikatorem jest *ranger* z liczbą drzew równą 50 wraz z selekcją funkcją *JMIM* z parametrem $k = 12$.
Otrzymane zmienne były takie same jak w przypadku funkcji *bagging*.

Prawdopodobieństwo dla zbioru testowego: 0.503
Prawdopodobieństwo dla zbioru walidacyjnego: 0.7466667

AUC dla zbioru testowego: 0.492
AUC dla zbioru walidacyjnego: 0,821

# Podsumowanie

- Wartości AUC nie różniły się mocno poza klasyfikatorem z danymi indywidualnymi i zbiorem testowym i walidacyjnym
- Dla zbioru testowego prawdopodobieństwo oraz AUC wychodziło znacznie niższe niż dla zbioru walidacyjnego
- Ogólnie w przypadku selekcji zmiennych prawdopodobieństwo oraz AUC jest niższe niż w przypadku modelu pełnego, bo jest trochę mniej informacji
- Dopasowanie modelu przy zmiennych wyselekcjonowanych jest znacznie szybsze niż dla modelu pełnego, dlatego często taki model jest lepszy
- Wybrałem klasyfikator z modelem opartym na funkcji *randomForest* z liczbą zmiennych równą 6

Dziękuję za uwagę!