Praca domowa 10

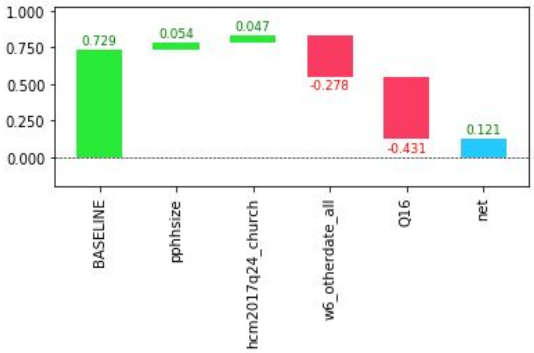
Ilona Bednarz

Dopasowany został model Random Forest przewidujący czy dana osoba jest zamężna na podstawie 4 zmiennych:

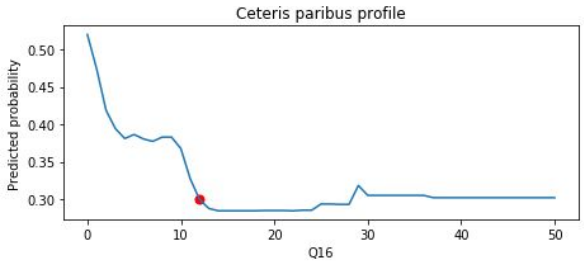
- Q16: How many of your relatives do you see in person at least once a month? (numeryczna)
- pphhsize: Household Size (numeryczna)
- hcm2017q24_church: met in or through church or religious organization (binarna)
- w6_otherdate_all: In past yr have you met anyone for dating romance or sex not incl current partner (zm. binarna)

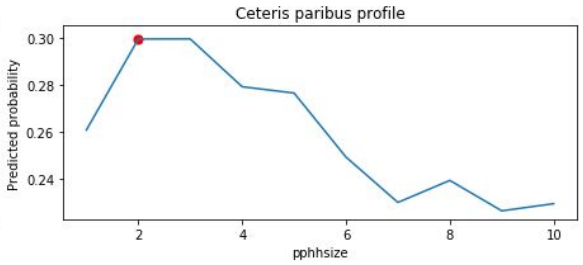
Wyjaśnienie działania modelu

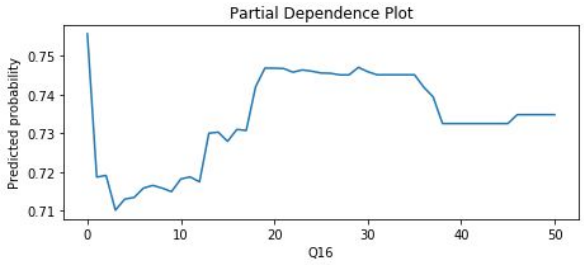
Omówmy zachowanie modelu dla przykładowej osoby ze zbioru danych: osoba ta spotkała 12 osób ze swojej rodziny przez ostatni miesiąc (zmienna Q16=12), mieszka w dwuosobowym gospodarstwie domowym (zmienna pphhsize=2), nie poznała swojego obecnego partnera na spotkaniu religijnym oraz w ciągu ostatniego roku była na randce z kimś innym niż obecny partner. Dla takiej osoby model stwierdza, że prawdopodobnie nie jest ona w związku małżeńskim (prawdopodobieństwo bycia w związku wynosi **0.3**). Zmienne pphhsize oraz hcm2017q24_church miały lekki dodatni wpływ na odpowiedź modelu (rysunek po prawej), natomiast w6_otherdate_all oraz Q16 zmniejszyły znacznie to prawdopodobieństwo.

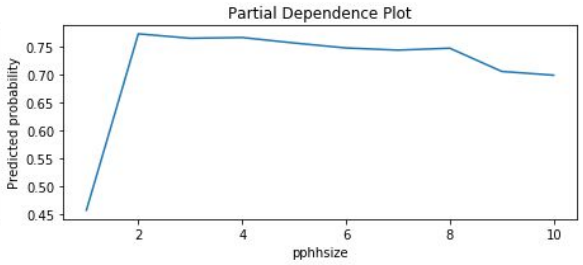


Na wykresach Ceteris Paribus widocznych obok możemy zobaczyć, jak zmieniłaby się predykcja dla rozważanej osoby, gdybyśmy zmieniali wartość jednej wybranej zmiennej ciągłej. Natomiast na wykresach PDP widzimy, jak średnio zmienia się predykcja modelu dla wszystkich osób w zbiorze danych w zależności od wartości wybranej zmiennej ciągłej. Zwróćmy uwagę, że zmienna Q16 ma bardzo mały wpływ na odpowiedź modelu - zakres w jakim zmienia się prawdopodobieństwo na wykresie PDP to (0.71, 0.75). Zmienna pphhsize wydaje się dużo bardziej istotna. W szczególności, dla pphhsize=1, co oznacza, że osoba mieszka sama, na wykresie PDP prawdopodobieństwo posiadania małżonka jest znacznie niższe niż dla pozostałych wartości, co jest zgodne z intuicją, gdyż osoby samotne zwykle mieszkają same. Zauważmy jednak, że na wykresie CP dla tej zmiennej zależność jest odmienna, tzn. minimum obserwujemy dla wartości powyżej 6. Oznacza to, że dla tej konkretnej osoby zmienna pphhsize ma inny wpływ niż średnio, a więc mogą istnieć jakieś korelacje między tą zmienną a jakąś inną.









Jak wyglądały wyniki dla podobnych osób?

Predykcje dla osób o parametrach podobnych do analizowanej wcześniej osoby są widoczne w tabeli po prawej stronie. Zwróćmy uwagę, że większość z nich jest bliska wcześniejszej predykcji wynoszącej 0.3, co dobrze świadczy o stabilności modelu. Predykcje odbiegające bardziej, to wiersze 1 i 4, gdzie znajduje się inna wartość zmiennej binarnej w6_otherdate_all. Oznacza to, że zmienna ta ma duży wpływ na predykcję w tym przypadku. Nie chodzenie na randki powoduje wzrost prawdopodobieństwa, że jest się w związku małżeńskim, co jest zależnością zgodną z intuicją. Natomiast zmiana wartości zmiennej binarnej hcm2017q24_church na przeciwną, jak również mała zmiana wartości obu zmiennych ciągłych nie zmieniają znacząco predykcji.

Q16	pphhsize	hcm2017q24_church	w6_otherdate_all	prediction
15	1	1	0	0.495560
9	1	1	1	0.207069
13	2	0	1	0.288330
11	1	0	0	0.541967
13	3	0	1	0.289840

Jak można podnieść skuteczność modelu?

Aby podnieść skuteczność modelu proponowałabym tuning parametrów lasu losowego. Można również zastąpić zmienne nieistotne jakimiś innymi lub włączyć do modelu wszystkie niewykorzystane kolumny ze zbioru danych i wtedy dokonać selekcji zmiennych.