Zadanie domowe 10

Wojciech Mańke

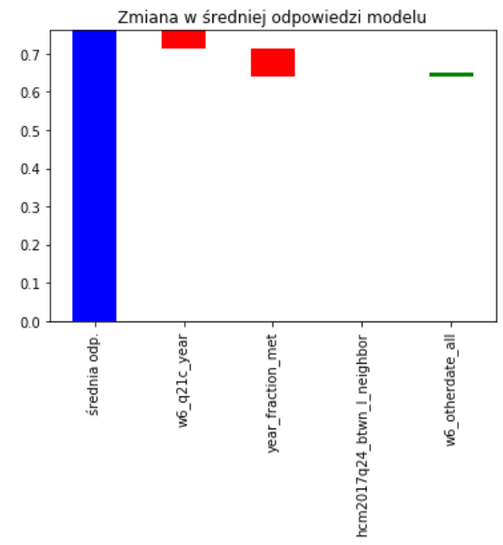
W oparciu o zbiór danych *hcmst2017* zbudowany został model XGBoost przewidujący, czy dana osoba jest żonata/zamężna. Model opiera się na następujących zmiennych:

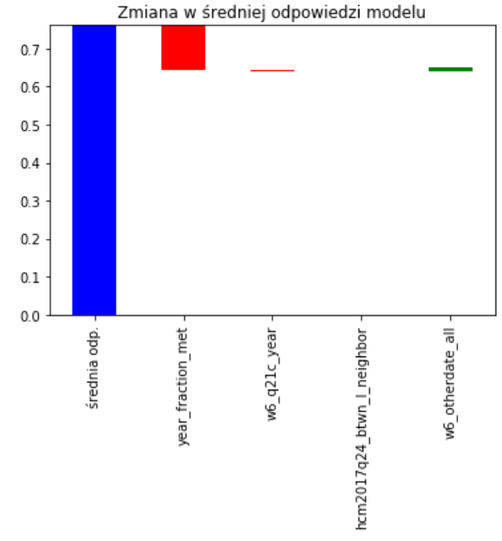
- *w6_q21c_year* - w którym roku po raz pierwszy para zamieszkała ze sobą,
- *year_fraction_met* - rok poznania, z uwzględnieniem miesiąca,
- *hcm2017q24_btwn_l_neighbor* - czy para była sąsiadami przed spotkaniem,
- *w6_otherdate_all* - czy w zeszłym roku dana osoba poznał(a) kogoś poza partnerem.

W dalszej części będę analizował osobę, która poznała swojego partnera w połowie **marca 2005 roku**. Para zamieszkała ze sobą na początku roku **2009.** Para **nie była sąsiadami** przed związkiem. Analizowana osoba **nie spotkała** nikogo w kontekście romantycznym w ciągu ostatniego roku. Według modelu prawdopodobieństwo, że dane osoby są w małżeństwie wynosi około **65%**.

Dlaczego?

Rozważmy wykresy pokazujące zmianę średniej odpowiedzi modelu dla dwóch wybranych kolejności zmiennych.

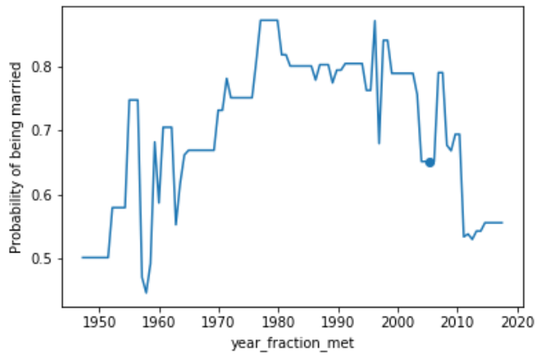


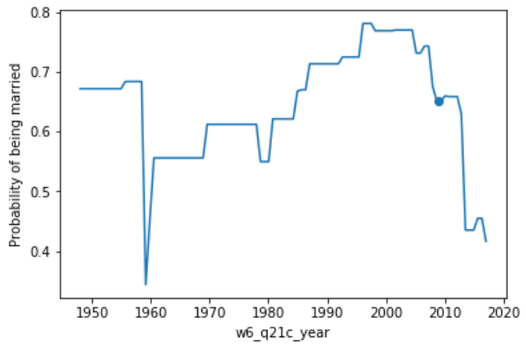


Jak widać zarówno rok poznania, jak i rok pierwszego zamieszkania ze sobą mają negatywny wpływ na prawdopodobieństwo małżeństwa. Dodatkowo wpływ roku zamieszkania jest minimalny pod warunkiem, że znamy już rok poznania.

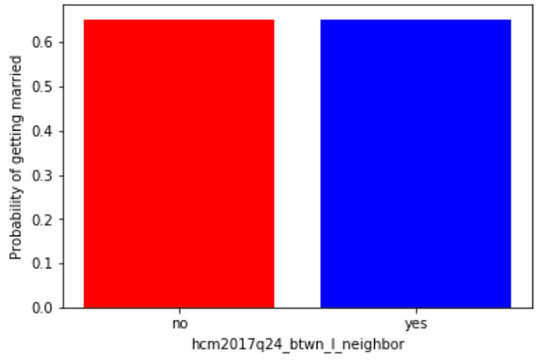
Co należałoby zmienić?

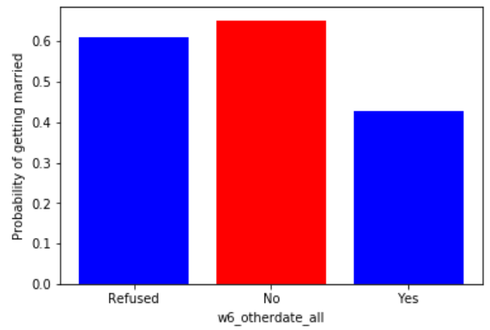
Przeanalizujmy profile Ceteris Paribus dla zmiennych dotyczących lat.





Wynika z nich, że prawdopodobieństwo, że analizowane osoby są małżeństwem wzrosłoby znacznie (do około 79%) gdyby para poznała się w 1999 r. Podobnie, gdyby te osoby zamieszkały razem w 2007 roku (a więc 2 lata po poznaniu) prawdopodobieństwo małżeństwa wynosiłoby 74%.
Wykresy dla zmiennych dyskretnych wskazują, że ich zmiana albo obniżyłaby predykcję, albo nie miałaby na nią wpływu (kolor czerwony na wykresie odpowiada rzeczywistej wartości wyjaśnianej obserwacji).

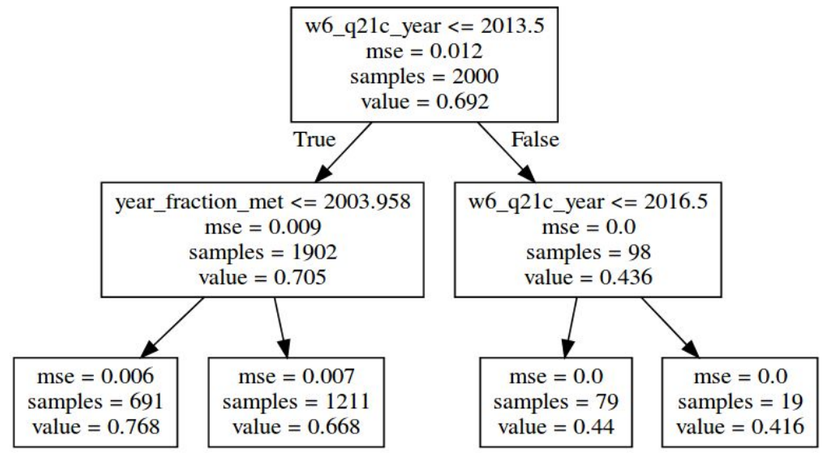




Analiza podobnych osób

Częściowa analiza podobnych osób (różniących się jedną zmienną) została wykonana w poprzednim punkcie przy okazji wykresów CP.

Dodatkowo możemy przeprowadzić lokalną analizę zachowania modelu za pomocą algorytmu LIME. Używając drzewa decyzyjnego jako wyjaśnienia otrzymujemy:



Na tej podstawie możemy dojść do podobnych wniosków jak poprzednio: zmniejszenie roku poznania oraz roku zamieszkania ze sobą podnosi prawdopodobieństwo małżeństwa. Podobnie zwiększanie wartości tych cech obniża to prawdopodobieństwo.