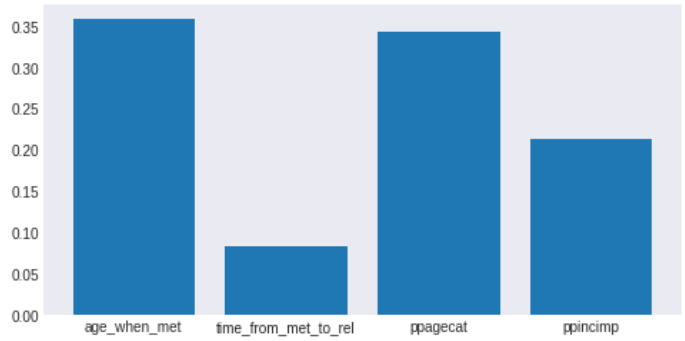
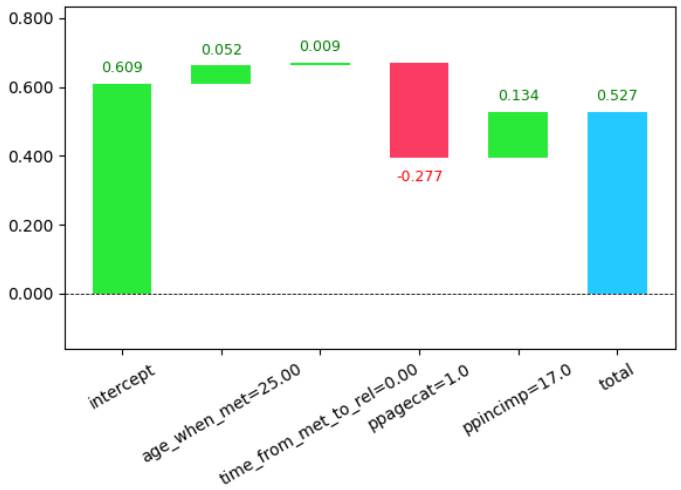
Bartosz Paszko PD10

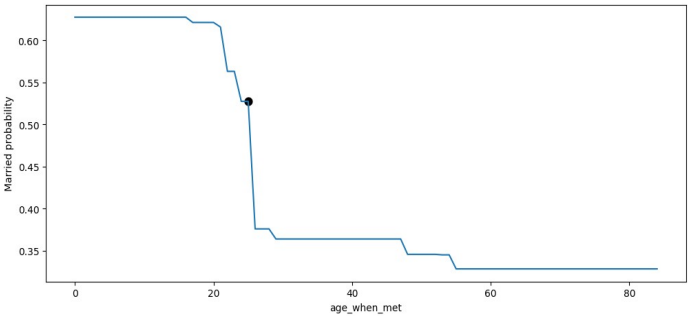
Na zbiorze *hcmst2017* został wytrenowany model XGBoost na następujących zmiennych:

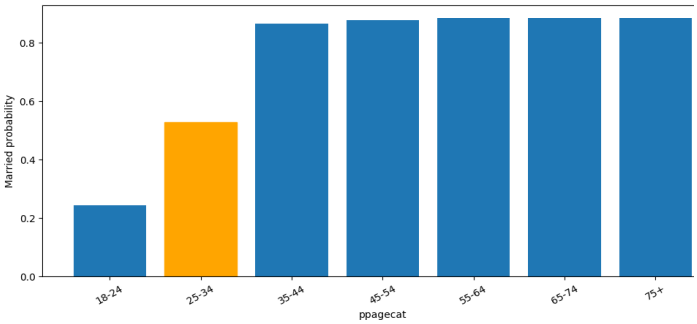
'age_when_met': wiek spotkania partnera – zmienna ilościowa
'time_from_met_to_rel': czas od spotkania partnera do związku – zmienna ilościowa
'ppagecat': wiek - zmienna kategoryczna
'ppincimp': przychód – zmienna kategoryczna



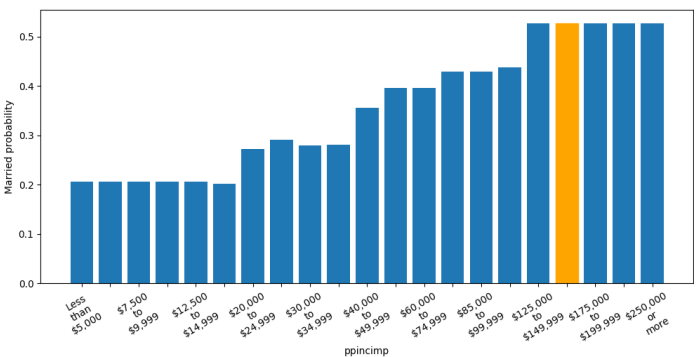
Zanim przejdziemy do analizy przykładów możemy sprawdzić jakie zmienne są istotne dla modelu. Na wykresie obok widać, że czas od spotkania do związku zdecydowanie mniej wpływa na predykcje. Z przeciwnej strony widać, że wiek ma bardzo duży wpływ poprzez obie związane z nim zmienne.
Dla wybranej losowo osoby (dla ułatwienia nazwijmy ją Jack) ze zbioru danych możemy przeprowadzić analizę odpowiedzi modelu. Jack opisany jest następującymi wartościami zmiennych (25, 0.0, 25-34, 150.000-174.999) w kolejności w jakiej zostały wymienione na początku. Odpowiedź modelu w tym przypadku to 0.527 co oznacza, że ma on ~53% szans na bycie w związku małżeńskim.
Na początku odpowiedzmy na pytanie **dlaczego** model tak twierdzi. Jak widać na wykresie BreakDown (2 od góry) średnia szansa bycia w związku małżeńskim wynosi aż 61%, za to młody wiek Jacka sprawia, że jego szanse spadają prawie o połowę. Jednak inne zmienne, a w szczególności przychód pomniejszają ten spadek ustanawiając ostateczny wynik 53%.
Możemy zapytać **co poprawiłoby jego szanse** na udany związek. Tutaj z pomocą przychodzą wykresy what-if.







Na wykresach obok widzimy jak zmieniłyby się szanse Jacka, gdyby zmodyfikować trochę jego dane. Otóż na pierwszym widać, że miałby dużo większe szansę na związek z partnerką, gdyby poznał ją przynajmniej 5 lat wcześniej. Drugi pokazuje, że Jack jest jeszcze młody i jego szanse wzrosną po przekroczeniu granicy 35 lat. Jeżeli chodzi o zarobki (3) to Jack zarabia wystarczająco, aby maksymalizować swoje szanse
Ostatnim elementem jest **porównanie odpowiedzi modelu** dla podobnych (pod względem rozpatrywanych cech) Jackowi osób. Wybierzmy 5 z nich do porównania (tabela poniżej). Jak widać prawdopodobieństwa są zbliżone do szans Jacka w 3 pzypadkach. Natomiast w 2 spadają o około 15%. Jest to zapeewne spowodowane ich wyższym wiekiem poznania partnera, co bardzo dobrze pokazuje 1 wykres po lewej.



age_when_met	time_from_met_to_rel	ppagecat	ppincimp	prediction
27.0	0.000000	1	17	0.375873
23.0	0.083252	1	17	0.563156
25.0	0.000000	1	18	0.527431
26.0	0.083374	1	16	0.375873
25.0	1.083374	1	17	0.523860