Podsumowanie i synteza wyjaśnień budowanych przez semestr.

W skrócie

Za pomocą ankiety **HCMST 2017**, wystawionej w lecie 2017 r., zebrano takie dane osobiste 3510 dorosłych **Amerykanów**, którzy posiadają lub posiadali życiowych partnerów, które miały posłużyć się do ustalenia czy osoba badana zawarła **związek małżeński** z partnerem, czy postanowiła się z nim rozstać.
Ten nowy zestaw danych jest dostępny na osobnej stronie https://data.stanford.edu/hcmst2017.
Na zebranym zbiorze danych wybrano **cztery cechy (predyktory)**, które miały posłużyć do predykcji statusu małżeńskiego (**zmienna predykowana**).
By to zrobić zbudowano na nim **klasyfikator** posługując się paradygmatem sztucznej inteligencji o nazwie „**uczenie maszynowe**”, który jako wynik, po zaaplikowaniu do tego modelu danych, zwraca prawdopodobieństwo bycia w związku.
W tym raporcie, analizuje odpowiedź zbudowanego modelu predykcyjnego dla przypadku wyszczególnionej osoby.

Wprowadzenie

Dane na których zbudowano model zawierały informację, po przetworzeniu, o **2744 badanych**.

Cecha Predykowana:
S - czy dana osoba poślubiła swojego partnera

Cechy predykcyjne, które zostały wybrane z zebranych danych to:
hcm2017q24_internet_other - czy osoba badana poznała partnera przez internet (Prawda/ Fałsz)
PPT01 – czy osoba badana mieszka z dziećmi (Prawda / Fałsz)
Q16 – Jak wielu krewnych osoba badana widywała w ciągu miesiąca osobiście (l. całkowita)
age_when_met – wiek w którym osoba badana poznała partnera (l. całkowita)

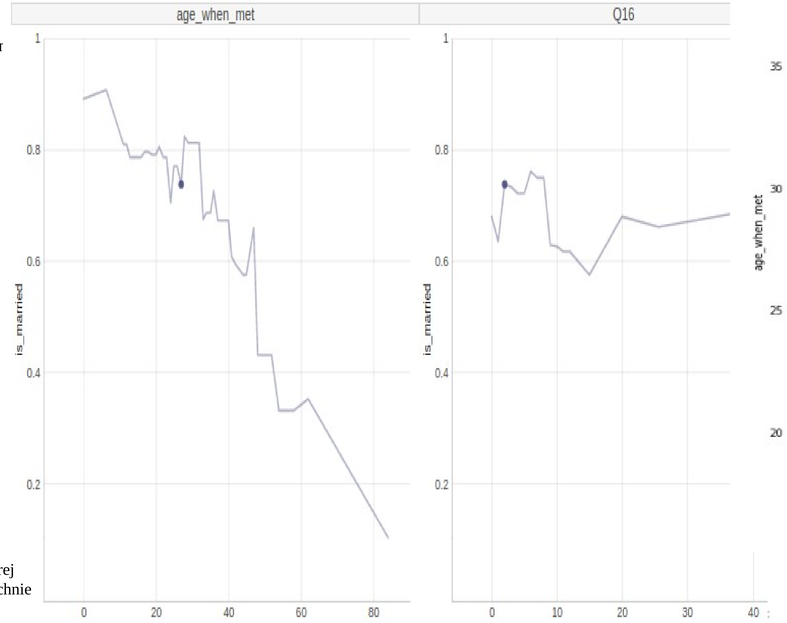
Zbudowany model oparty jest na architekturze **Extreme Gradient Boosting**, której implementacje wybrano **w języku Python** używając **biblioteki xgboost**, powszechnie znanej w społeczności uczenia maszynowego.
Dane zostały podzielone na treningowe oraz testowe, gdzie model został wytrenowany na pierwszy, a jego jakość oceniono za pomocą **miary AUC** – pola pod **krzywą ROC**. Ten wynik wynosi **0.67** po zaokrągleniu do 3 miejsc po przecinku.

Wybrana osoba ze zbioru testowego co do której zostanie przeanalizowana odpowiedź modelu to:
Osoba nie mieszkająca z dziećmi, spotykająca partnera w wieku 27 lat nie przez internet, która widzi się z dwójką krewnych w ciągu miesiąca
Znajduje się ona w prawym odchyleniu ćwiartkowym dla obu zmiennych **Q16** i **age_when_met**, przy czym dla obu zmiennych odpowiadające im wartości leżą poniżej średnich otrzymanych z danych.

Wytrenowany model dla wybranej osoby zwrócił prawdopodobieństwo **0.74** bycia w związku małżeńskim. Co jest bliskie danym z ankiety, które wskazują, że osoba ta rzeczywiście jest w związku małżeńskim

Okolice obserwacji

Wykresy **Ceteris Paribus** pozwalają nam się przyjrzeć zmianie odpowiedzi modelu, przy zmianie dokładnie jednego predyktora, podczas gdy inne predyktory mają ustalone wartości. Na dole widzimy takie wykresy dla zmiennych porządkowych **age_when_met** oraz **Q16**.

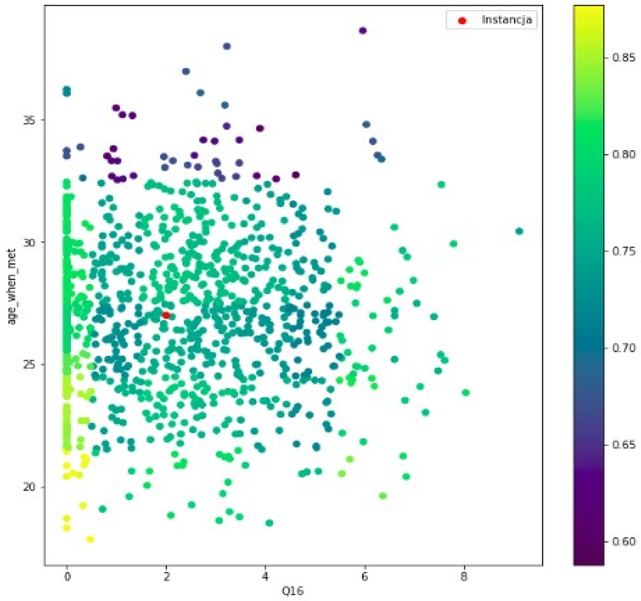


Z wykresów możemy zaobserwować iż:

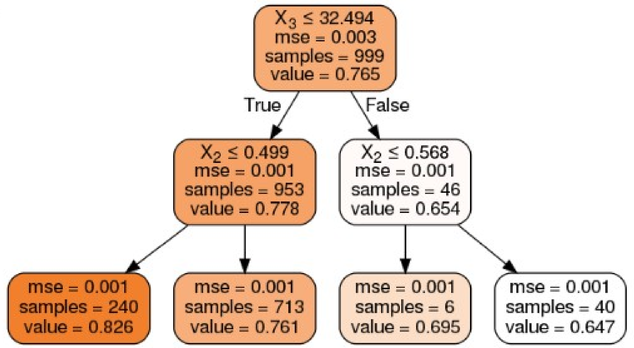
a) trend dla zmiennej **age_when_met** jest spadkowy i wręcz liniowy, twierdząc że prawdopodobieństwo dla tej osoby byłoby największe przy spotkaniu partnera w wieku przedszkolnym, ale im później by go spotkała, tym mniejsze byłoby prawdopodobieństwo zaślubin z nim, co jest sprzeczne z moją intuicją, że osoby młodsze mogą pozwolić sobie na większy wybór, który pozwala im być wybredniejszymi.

b) trend dla drugiej zmiennej **Q16** niewiele odstaje od funkcji stałej, oprócz sinusoidalnego pełnego cyklu na przedziale **[0;20]**, o wychyleniu **0.05**. Tu intuicja podpowiada mi, że osoby unikające bliskich jak i zżyte z nimi nazbyt, mogą mieć problemy z utrzymaniem związku, z powodu samolubności, czy braku samodzielności

Przybliżając odpowiedź modelu, na okolicę obserwacji, modelem z wysoką interpretowalnością, takim jak na przykład **drzewo decyzyjne**, możemy dowiedzieć się jak zachowuje się nasz model w okolicach wybranej instancji.

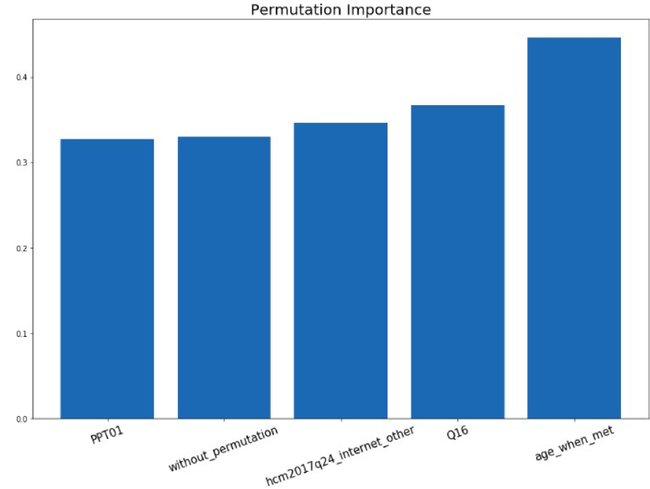


Co pozwala nam dostać lokalny system decyzyjny przybliżający odpowiedź modelu (**X3 – age_when_met**, **X2 -Q16**)



Globalny wgląd

Ostatnie drzewo decyzyjne wskazało na lokalnie największą istotność cechy **age_when_met**, zbadajmy czy jest tak w ogólności.



Badanie wkładu każdej zmiennej, przy spermutowaniu jej przy ustalonym porządku innych, względem miary **AUC** wskazuje, iż zmienna **age_when_met** jest najistotniejsza dla modelu , podczas gdy wkłady innych niewiele się od siebie różnią, jak i od wyniku modelu bez zastosowanej permutacji. Co więcej widać, iż zmienna **PPT01** pogarsza nieznacznie wynik, z czego wynika, iż model powinien przyłożyć do tej zmiennej więcej uwagi. Pokazane wyniki wyliczone są dla zbioru testowego, po ówczesnym wytrenowaniu modelu na zbiorze treningowym.

Jak można poprawić wyniki modelu

Przede wszystkim należy zbadać wyniki modelu pod względem optymalizacji hiperparametrów, gdyż w tym przypadku zostały wybrane domyślne. Należy również zadbać o niezbalansowanie zebranego zbioru, ponieważ liczba osób po przynajmniej cywilnym „tak”, stoi do tych, bez tego doświadczenia, w stosunku **3:1**.
Oczywistym krokiem, jest próba przeszukania wszystkich dostępnych cech w zebranym zbiorze **HCMST 2017,** by wybrać te które najlepiej wyjaśniają cechę **S.**
Należy również zbadać korelacje między wybranymi zmiennymi, by pozbyć się tych skorelowanych z innymi które najmniej wnoszą do predykcji.