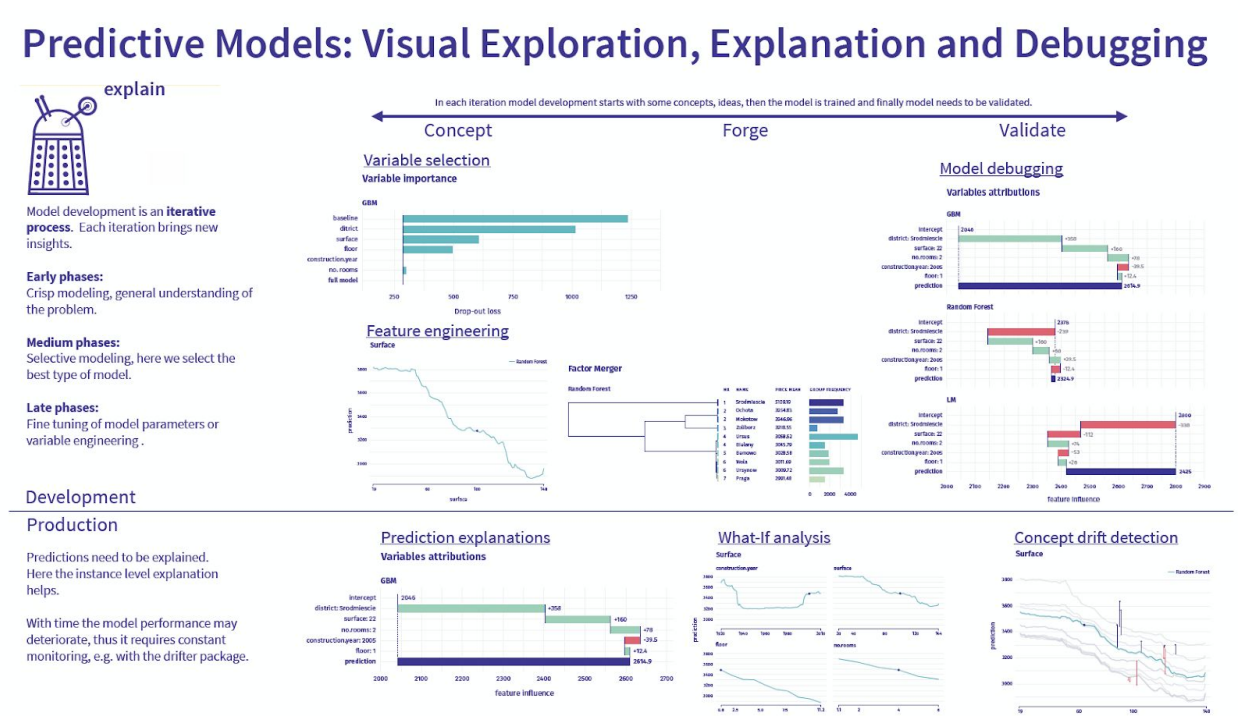


Wyjaśnialne Uczenie Maszynowe - Grupa 6.

07.06.2019

Imię i nazwisko

Artem Gerashchenko

Daria Hubernatorova

Dawid Lipski

Wojciech Mańke

Joanna Piega

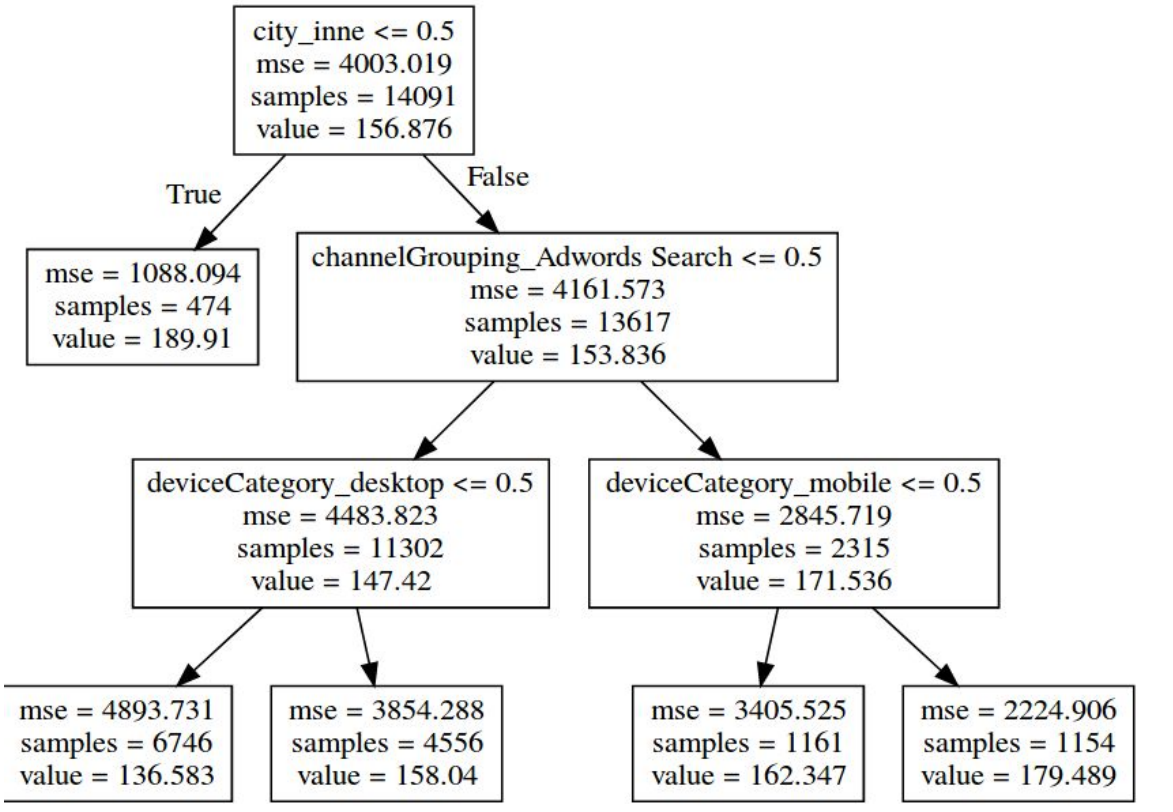
Cele

1. Przedstawienie ogólnej koncepcji projektu.
2. Wyjaśnienie modelu na poziomie globalnym.
3. Wyjaśnienie modelu na poziomie pojedynczej obserwacji.
4. Dodatkowe zagadnienia związane z projektem.

Rozwiązanie

I. Przedstawienie ogólnej koncepcji projektu.

W naszym rozwiązaniu zastosowaliśmy połączenie dwóch modeli. Dane zostały najpierw sklastrowane z użyciem drzewa decyzyjnego, a następnie dla każdego z klastrów została użyta regresja logistyczna przewidująca ilość kliknięć w dany produkt.



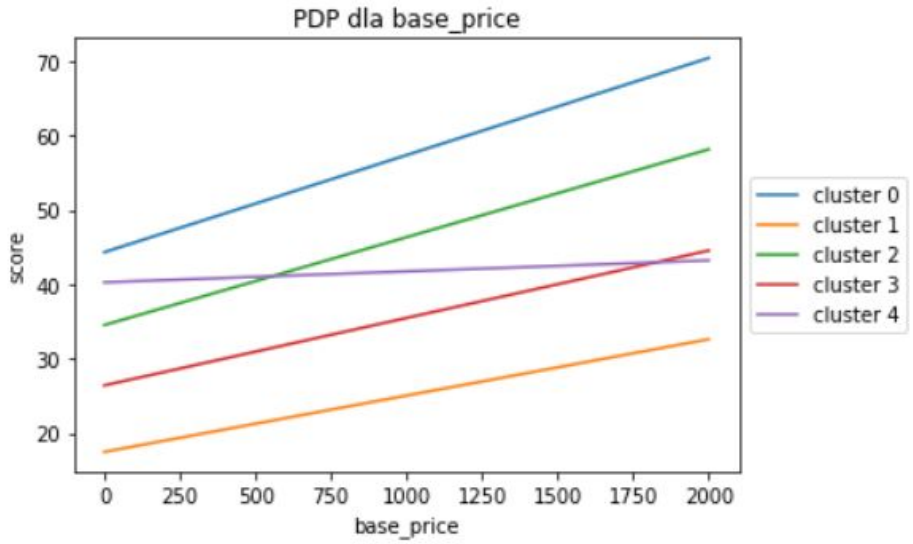
Zdecydowaliśmy się zastosować również prostą inżynierię cech: usunęliśmy niepotrzebne kategorie, wyczyściliśmy nazwy marek (usuwanie znaków inne niż litery) oraz podzieliliśmy miasta na nie-wojewódzkie (inne) oraz wojewódzkie. W

ostatecznej wersji modelu zrezygnowaliśmy z używania marek, gdyż nie dawały one dużej poprawy rezultatu, a znacząco zwiększały powstające formuły SQL.

II. Wyjaśnienie modelu na poziomie globalnym

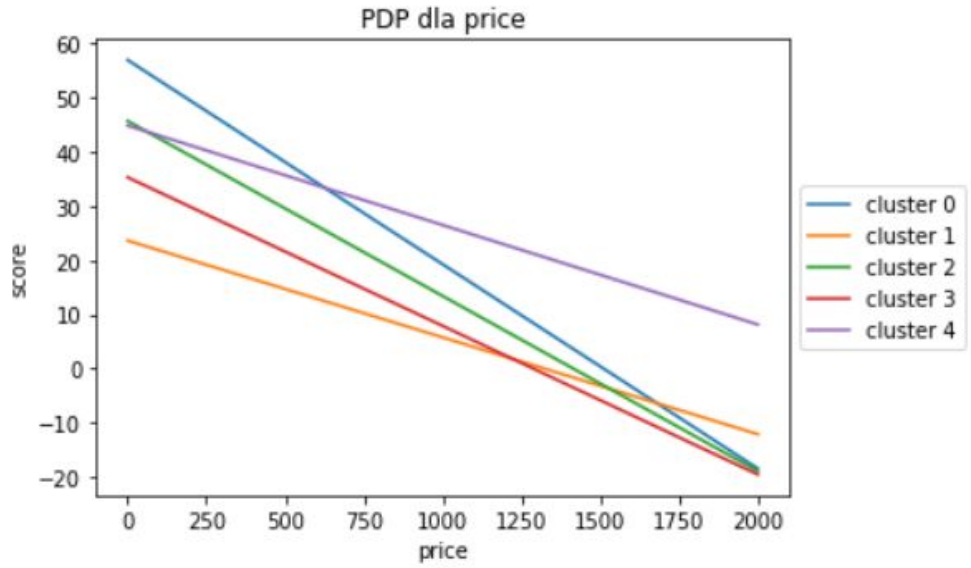
Regresja logistyczna, która została użyta jako model scorujący produkty wewnątrz klastrów jest modelem łatwym w interpretacji. Dzięki estymowanym współczynnikom możemy w prosty sposób ocenić charakter danej zmiennej w modelu. Jednakże patrząc na model globalnie - chcielibyśmy stwierdzić jaka jest różnica w działaniu modelu na poszczególnych grupach klientów. W tym celu wykreśliliśmy krzywe PDP dla istotnych zmiennych ciągłych.

1. W przybliżeniu pozioma linia dla base_price w klastrze 4 sugeruje, że dla klientów mieszkających w miastach wojewódzkich, nie dostających się na stronę przez reklamę, ani nie korzystających z telefonu - bazowa cena produktu nie wpływa na ilość kliknięć. Ciekawą obserwacją jest globalny trend rosnący, tzn. im wyższa bazowa cena produktu, tym więcej kliknięć. Mogłoby się to wydawać nieintuicyjne, jednakże spójrzmy na analogiczny wykres ceny.



2. Po przeanalizowaniu wykresu PDP dla ceny obserwujemy globalnie malejący trend. Oznacza to, że im wyższa jest aktualna cena produktu, tym mniej kliknięć. Porównując te krzywe z PDP dla ceny bazowej, można wyciągnąć wniosek, że klienci chętniej klikają w produkty, których bazowa cena jest wysoka, zaś aktualna niska, co świadczy o znacznej przecenie produktu. Obserwacja ta jest zgodna z intuicją. Potwierdza to dobre działanie modelu. Na poniższym wykresie, klienci z klastra 4 ponownie odstają od reszty. Można zauważyć, że są bardziej wrażliwi na wartość aktualnej ceny, niż ceny

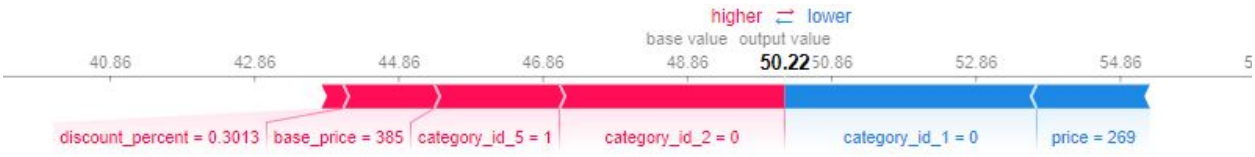
bazowej, gdyż krzywa odpowiadająca za klaster 4 ma większe nachylenie niż analogiczna krzywa na poprzednim wykresie.

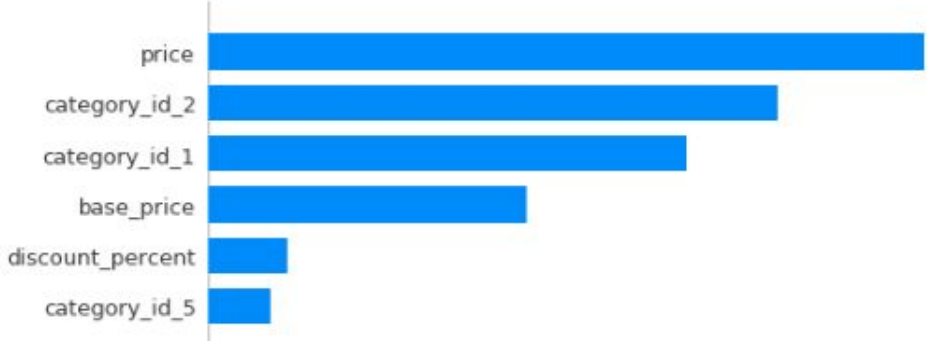


III. Wyjaśnienie modelu na poziomie pojedynczej obserwacji.

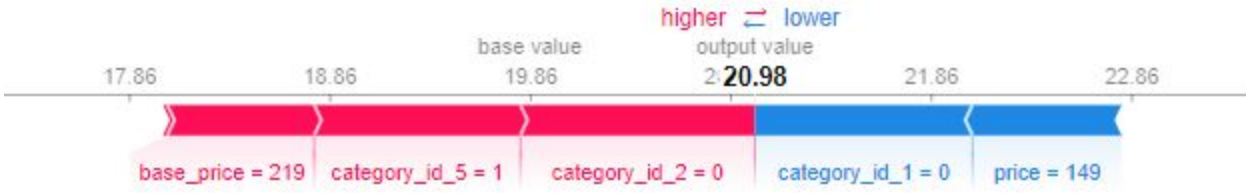
W tym etapie naszym celem było wyjaśnienie przedstawionego przez nas modelu na podstawie analizy pojedynczej obserwacji. Ze względu na hierarchiczność naszego modelu oraz prosty model na końcu procesu (liniowa regresja) zdecydowaliśmy się zrobić analizę opierając się na SHAP (SHapley Additive exPlanation). Dla każdego clustra (i modelu) zrobiono analizę dla losowo wybranej obserwacji:

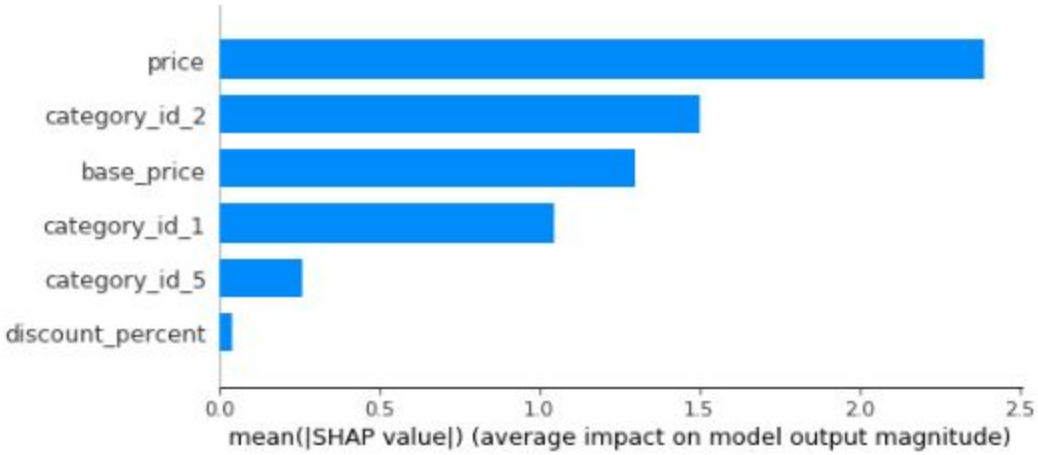
1) model 0:



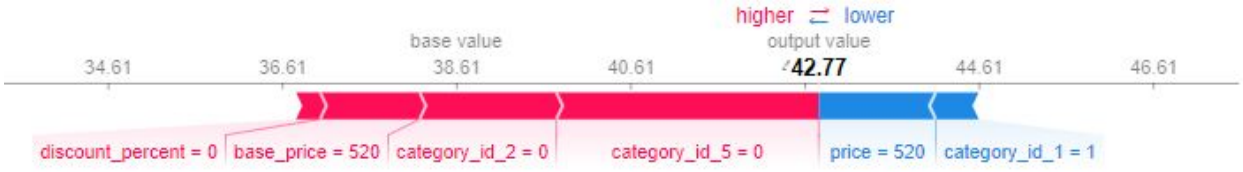


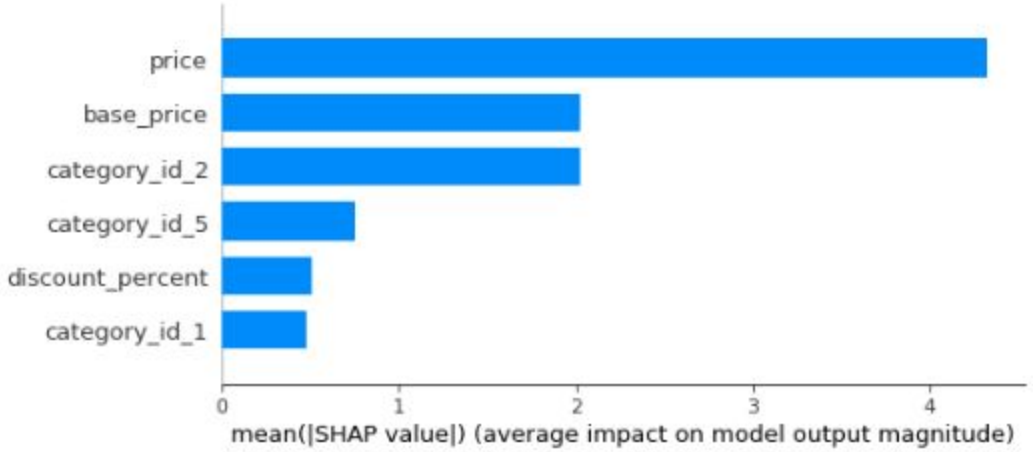
2) model 1:



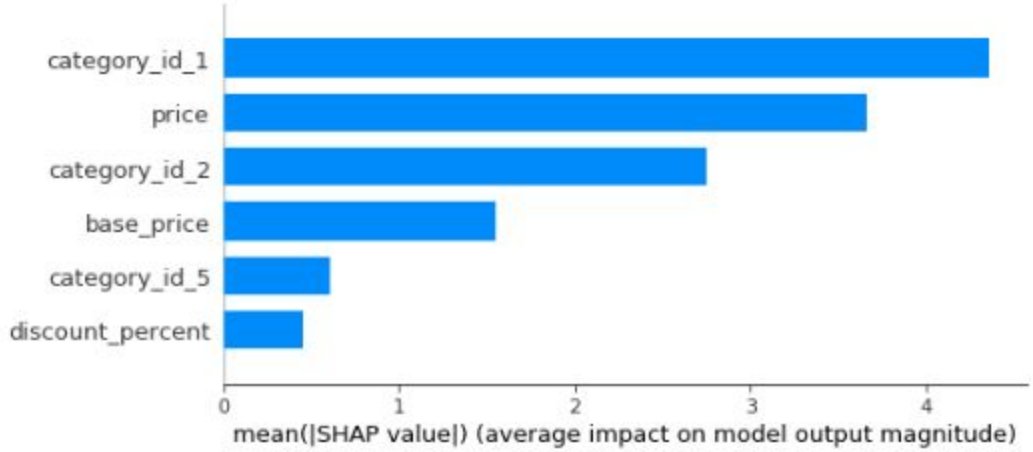


3) model 2:

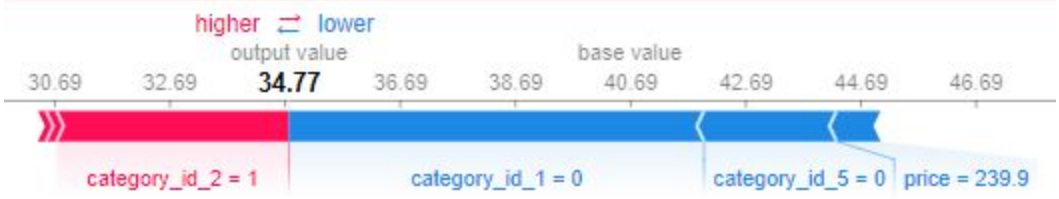


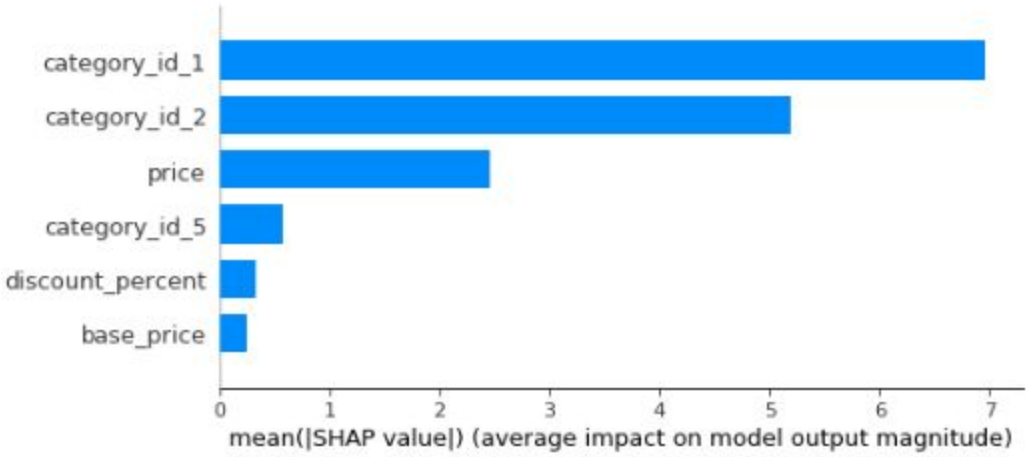


4) model 3:



4) model 4:





Jak można zauważyć, w zależności od klastra - a co za tym idzie modelu - różne zmienne wpływają na losową obserwację w różnym stopniu. Dla pierwszy klastrów wyraźnie widać, że to cena jest zmienną, która dominuje we wpływie na zmienną objaśnianą. W tych przypadkach powodem zazwyczaj jest zbyt wysoka cena danego produktu. Zauważono dość sporą zależność między ceną produktu a częstością kliknięć. Klienci nastawieni byli jednak na tańsze produkty, przez co zbyt duża cena dla danej obserwacji powodowała zmniejszenie liczby kliknięć. W ostatnich klastrach sytuacja zmieniła się nieznacznie, ale dalej cena kreowała dość sporą zmianę, choć w dwóch ostatnich to zmienna określająca, czy produkt jest ubraniem miała największe znaczenie.

IV. Dodatkowe zagadnienia związane z projektem.