Podsumowanie projektu - Wyjaśnialne Uczenie Maszynowe

Grupa III

Wojciech Celej, Daria Ilina, Zuzanna Opała, Olaf Skrabacz,
Julia Tonkiewicz, Piotr Wawrzyniak, Mateusz Zakrzewski

Pomysł:

Stwórzmy *super model* wykorzystujący wszystkie parametry i podzielmy go na dwa mniejsze modele

Główne problemy

Co właściwie ma przewidywać ten model?

- naturalne wydają się kliknięcia lub inne tego rodzaje wskaźniki
- **nowe danych**: wiemy tylko co było kliknięte, a nie wiemy zupełnie czemu
- **stare dane**: nieaktualne marki i generalnie nie do końca łączące się kolumny, ale dużo więcej informacji, na przykład o nieklikniętych produktach

Jak uprościć problem do prostej struktury kilku zmiennych i SQL?

- stworzenie dobrego rankingu wymaga raczej bardziej skomplikowanej struktury
- w części modelu oceniającej produkt możemy korzystać tylko ze zmiennych mało różnicujących: *kategoria, marka, cena*

Główne problemy

Co właściwie ma przewidywać ten model?

- naturalne wydają się kliknięcia lub inne tego rodzaje wskaźniki
- ~~**nowe danych**: wiemy tylko co było kliknięte, a nie wiemy zupełnie czemu~~
- **stare dane**: nieaktualne marki i generalnie nie do końca łączące się kolumny, ale dużo więcej informacji, na przykład o nieklikniętych produktach

Jak uprościć problem do prostej struktury kilku zmiennych i SQL?

- stworzenie dobrego rankingu wymaga raczej bardziej skomplikowanej struktury
- w części modelu oceniającej produkt możemy korzystać tylko ze zmiennych mało różnicujących: *kategoria, marka, cena*

Główne problemy

Co właściwie ma przewidywać ten model?

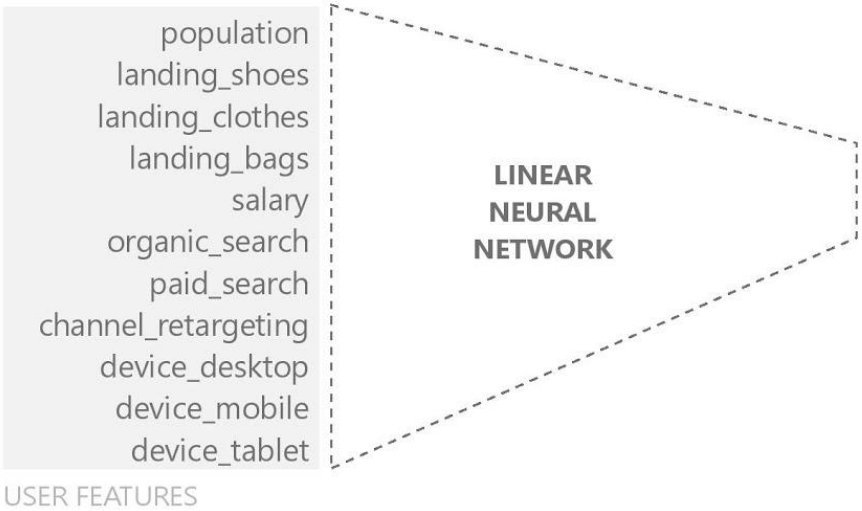
- naturalne wydają się kliknięcia lub inne tego rodzaje wskaźniki
- ~~**nowe danych**: wiemy tylko co było kliknięte, a nie wiemy zupełnie czemu~~
- **stare dane**: nieaktualne marki i generalnie nie do końca łączące się kolumny, ale dużo więcej informacji, na przykład o nieklikniętych produktach

Jak uprościć problem do prostej struktury kilku zmiennych i SQL?

- stworzenie dobrego rankingu wymaga raczej bardziej skomplikowanej struktury
- w części modelu oceniającej produkt możemy korzystać tylko ze zmiennych mało różnicujących: *kategoria, marka, cena*

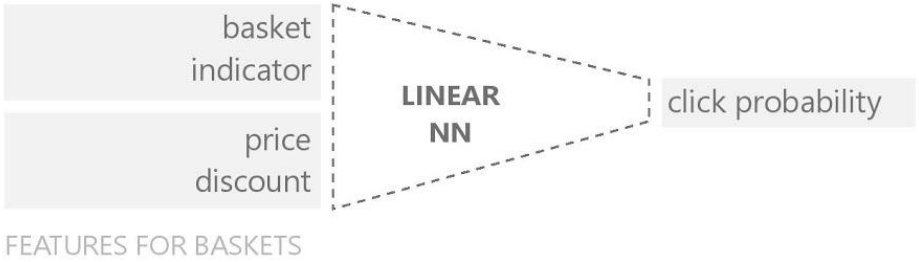
Bazowy model

Klasyfikacja użytkownika



Prawdopodobieństwo kliknięcia produktu

Basket - wskaźnik należący do przedziału <0, 1>;
Sieć uczy się kompresować dane o użytkowniku w taki sposób aby ten wskaźnik dawał możliwie dużo podczas predykcji - w przypadku tej sieci intuicja zgadzała się z uzyskanymi wynikami. Sieć osiągnęła identyczne rezultaty co metody wykorzystujące do predykcji wszystkie parametry.

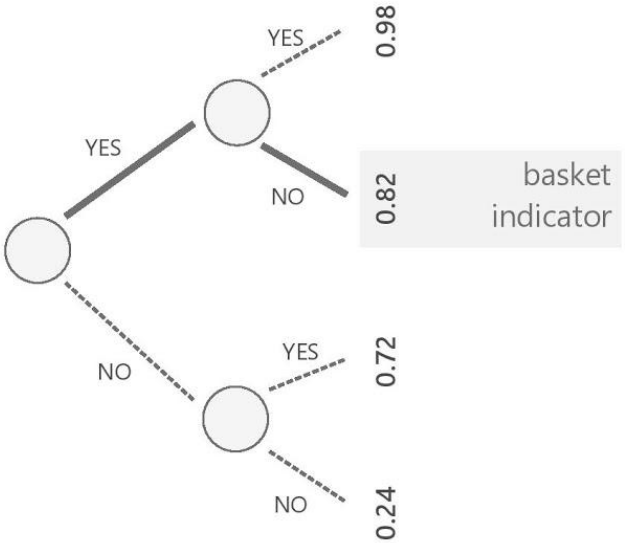


Model osiągnął takie same wyniki co model wykorzystujący wszystkie parametry do predykcji

Model zgodny z wymaganiami

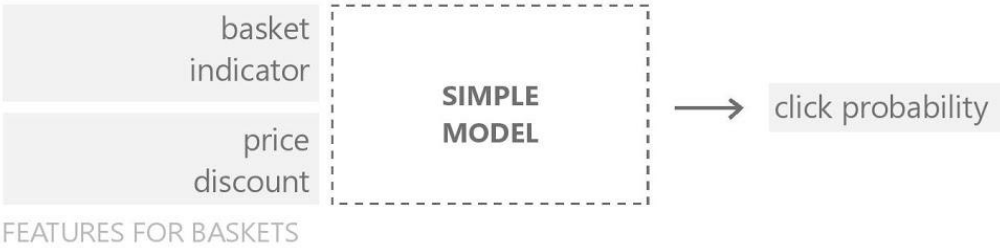
Klasyfikacja użytkownika - aproksymacja

Aproksymacja cechy "basket" wyznaczonej przez sieć neuronową za pomocą drzewa.



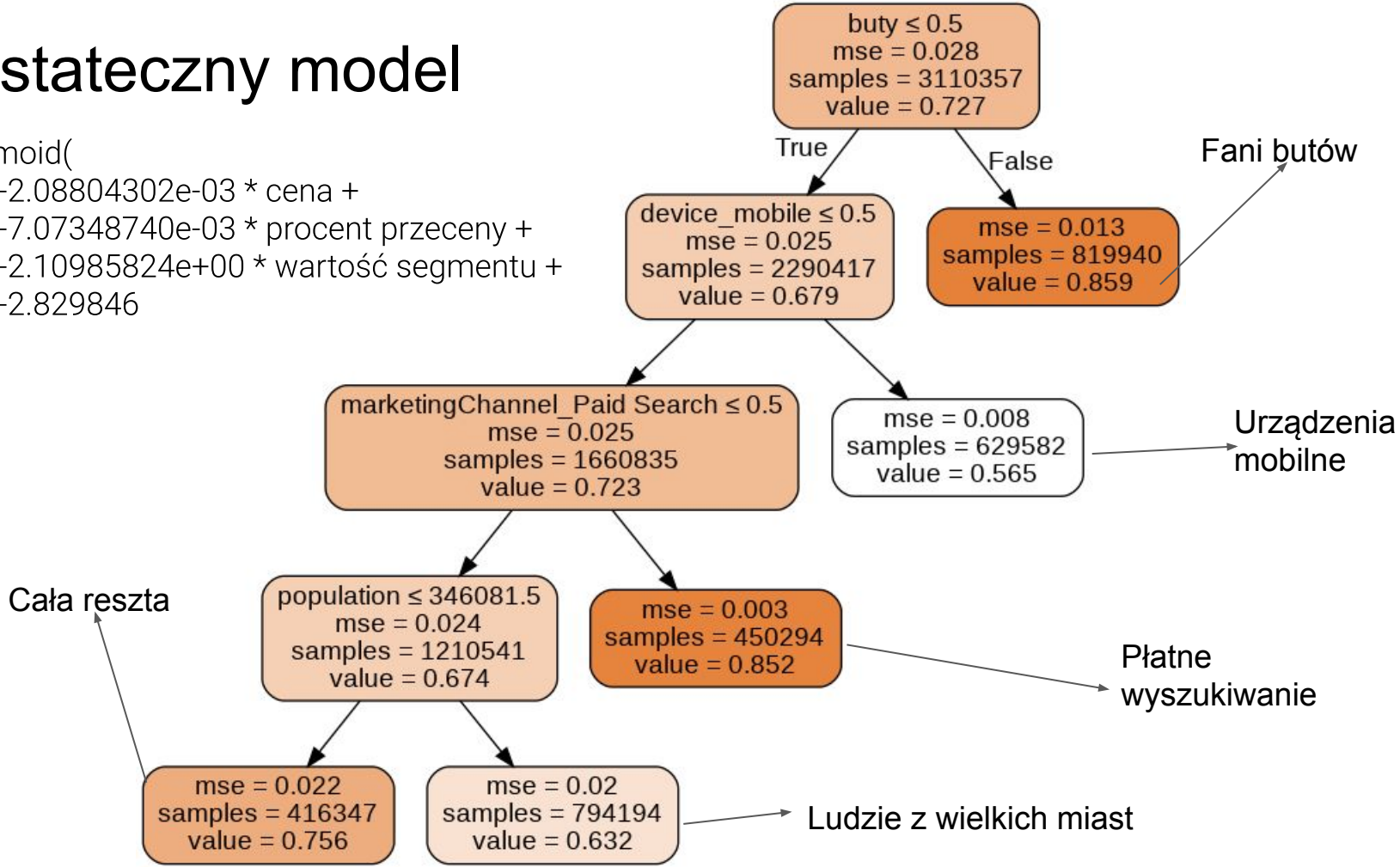
Prawdopodobieństwo kliknięcia produktu

Bazując na stałym współczynniku "basket" tworzymy prosty model



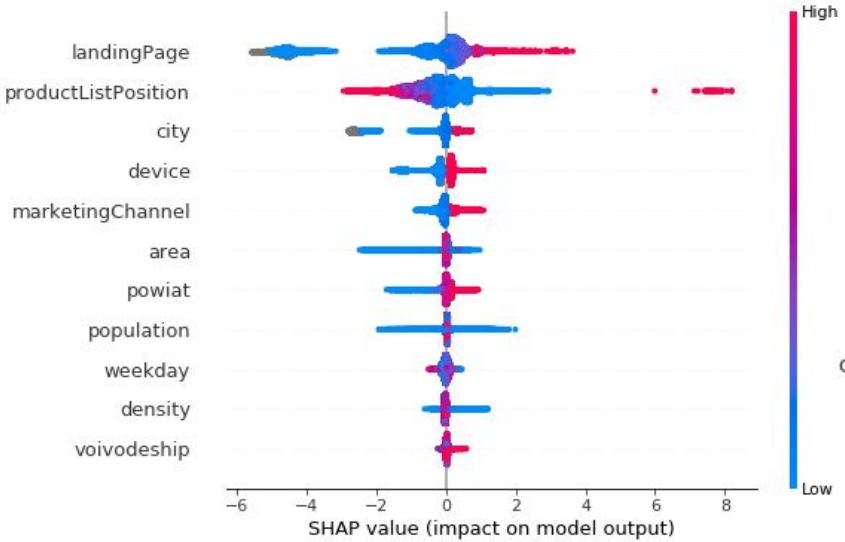
Ostateczny model

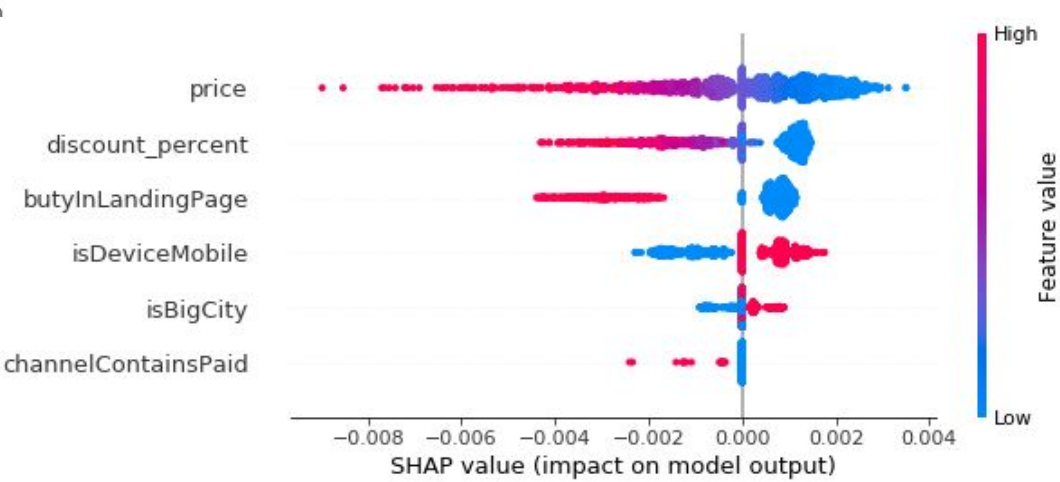
sigmoid(
-2.08804302e-03 * cena +
-7.07348740e-03 * procent przeceny +
-2.10985824e+00 * wartość segmentu +
-2.829846
)



Wyjaśnienie modelu

SHAP Values

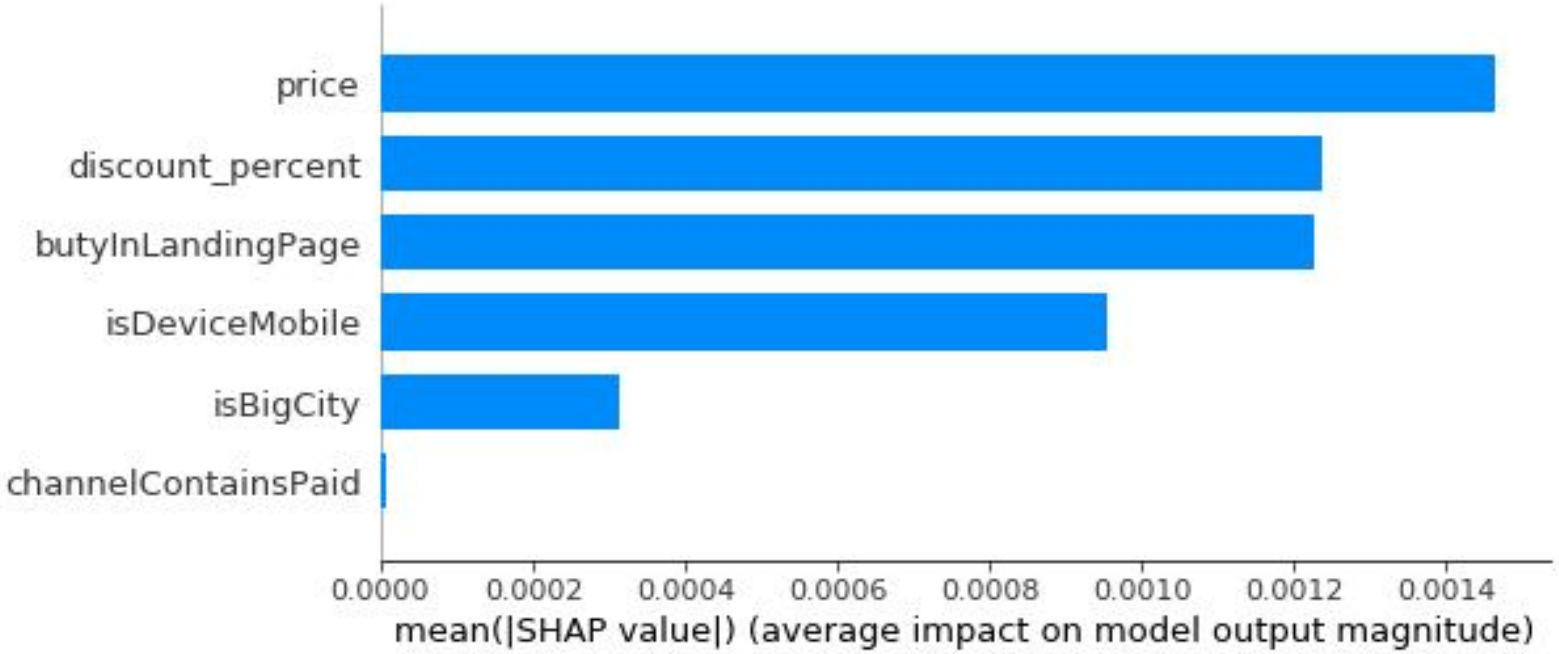




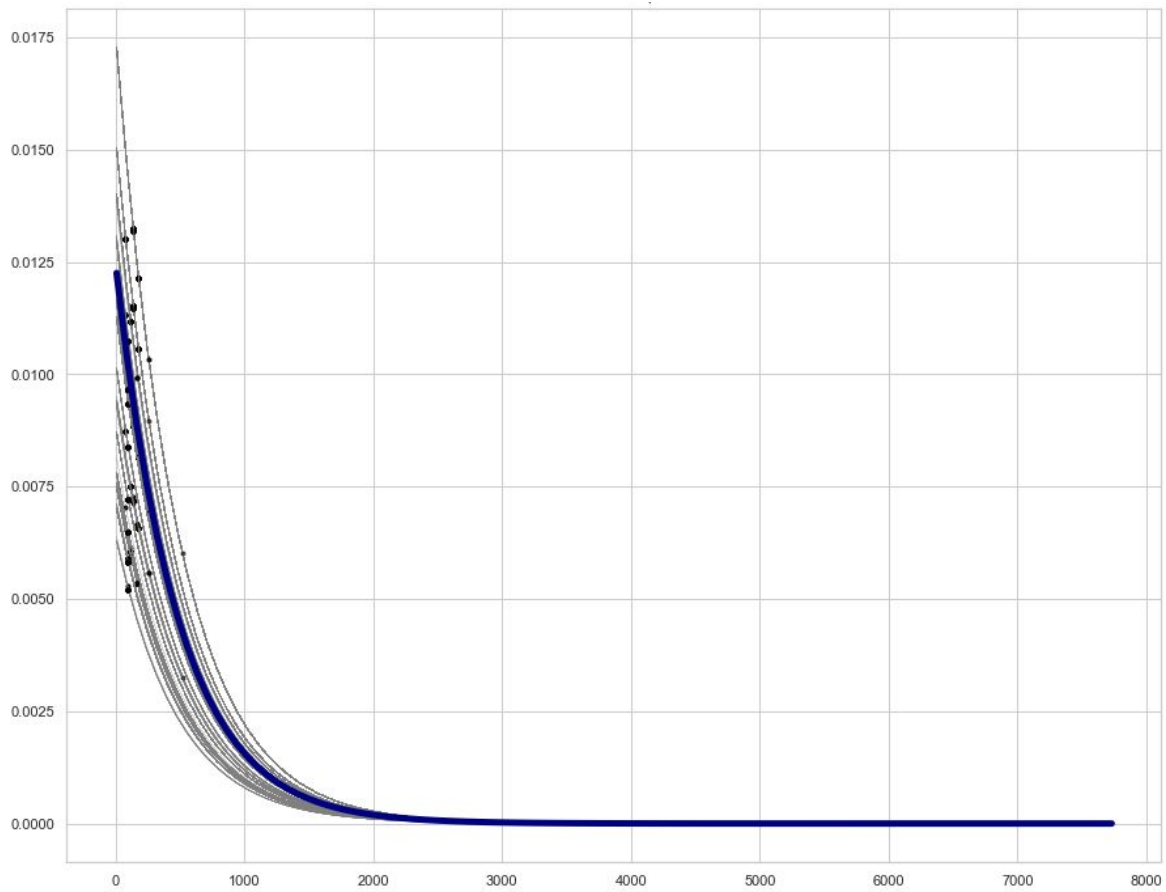
stare podejście

nowe podejście

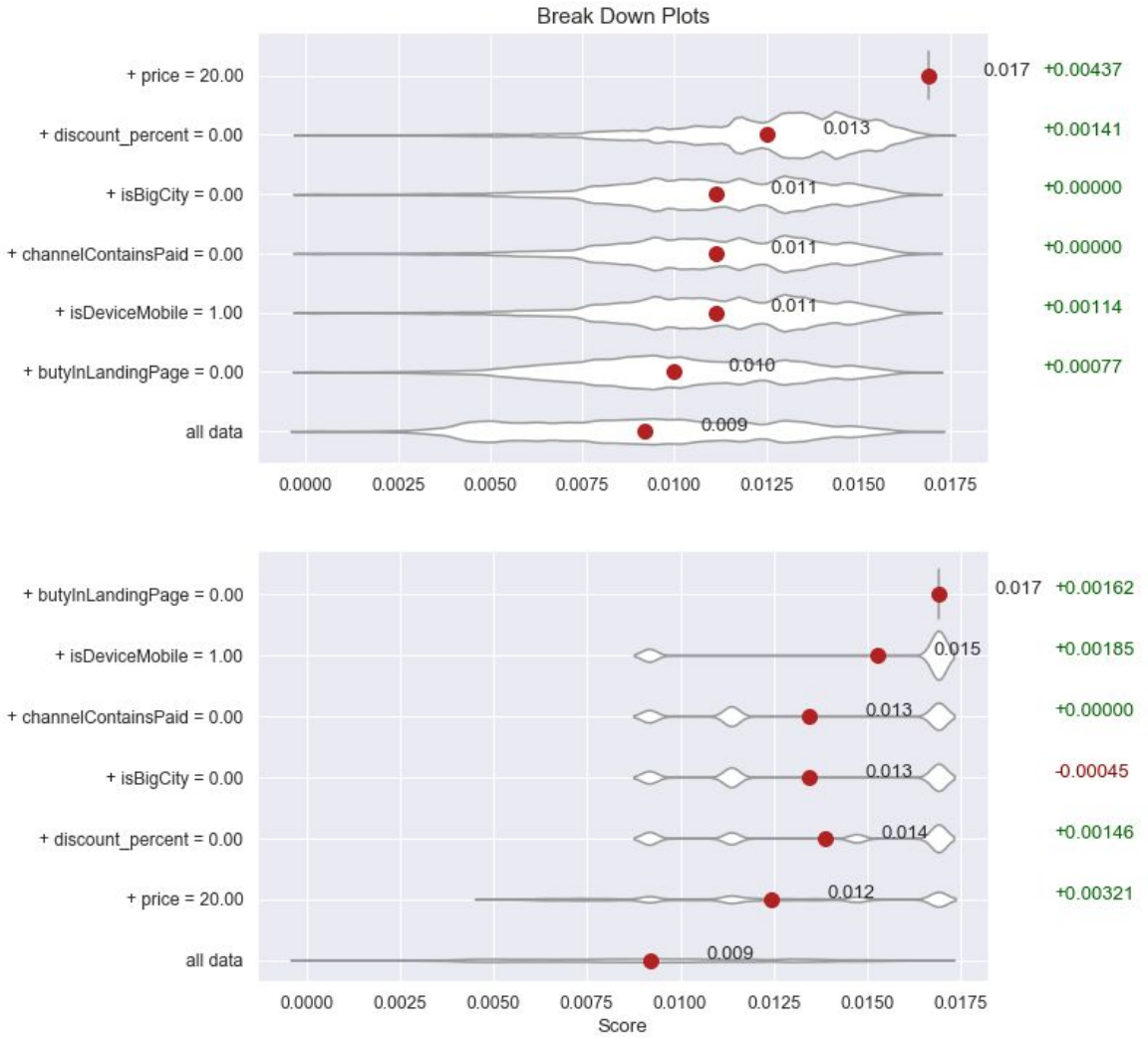
Istotność zmiennych na podstawie SHAP



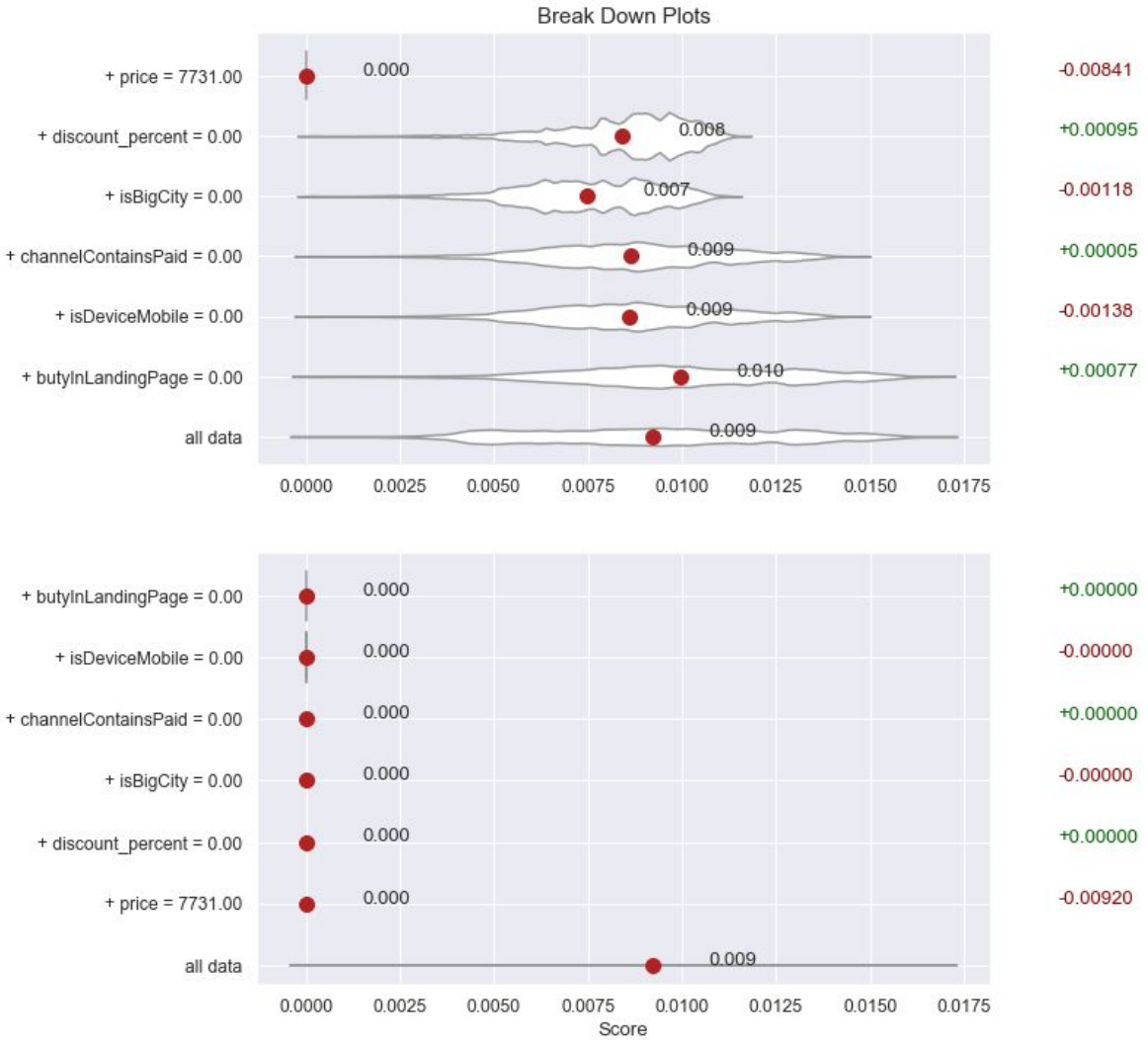
Wykres PDP + CP dla zmiennej *price*



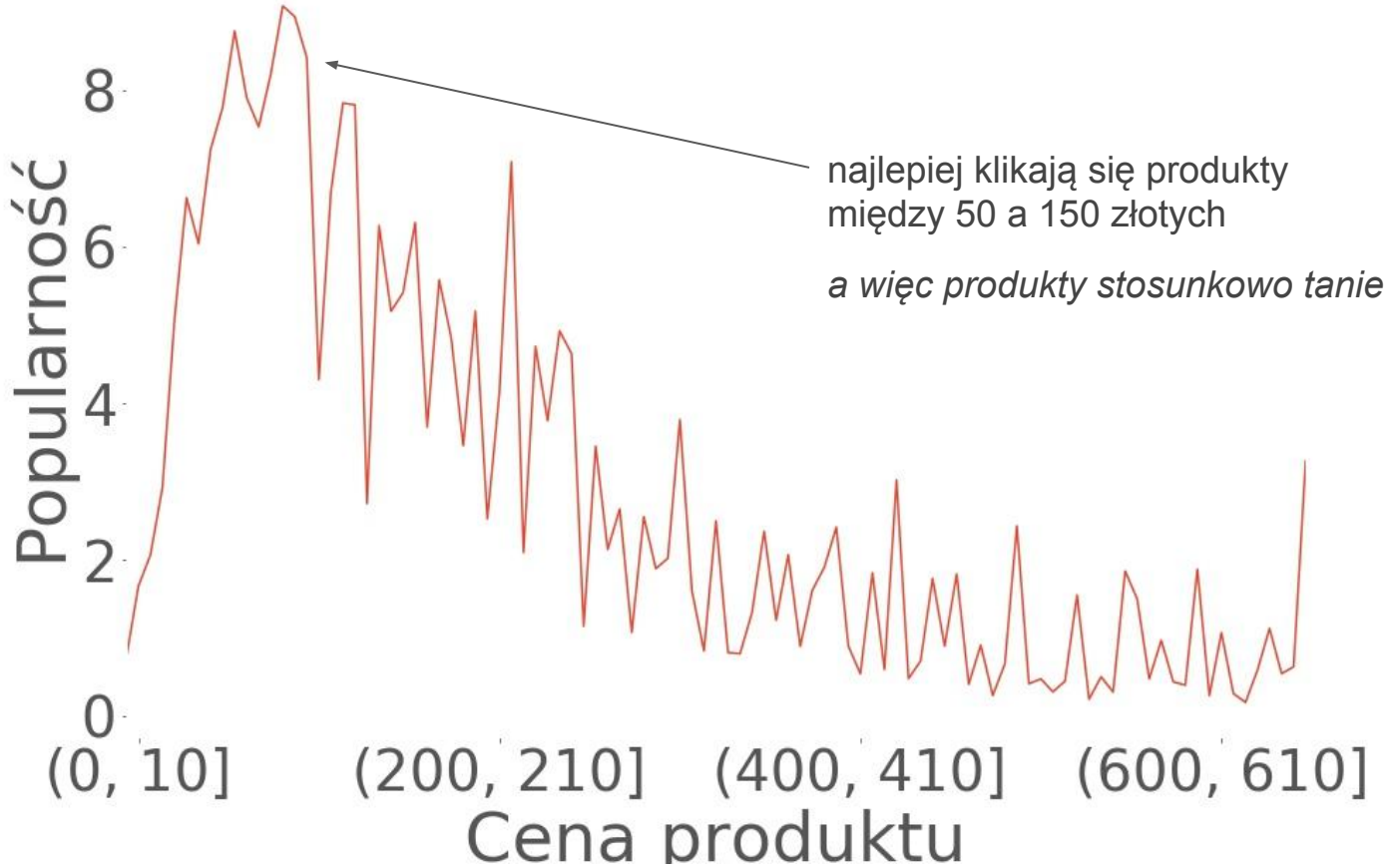
Wykresy Break Down dla najwyżej ocenionej próbki



Wykresy Break Down dla najniżej ocenionej próbki



Stosunek kliknięć do liczby ofert w danej cenie



Podsumowanie

- nasz model składał się z dwóch części różniących się jedną wartością dla poszczególnych koszyków
- użyliśmy starych danych które pozwalały przewidywać kliknięcia
- głównym aspektem, na który patrzył model była cena i procent przeceny

Co mogliśmy zrobić lepiej

- stworzyć bardziej skomplikowany model niż liniowy
- jakoś użyć informacji o marce (może encoding na innym zadaniu jako, że informacji o klikaniu w nowych danych nie było)