

# Zastosowanie uczenia maszynowego w praktyce

Data Science in Practice x LaModa

07.06.2019

Zespół 2

Tomasz Klonecki

Bartosz Paszko

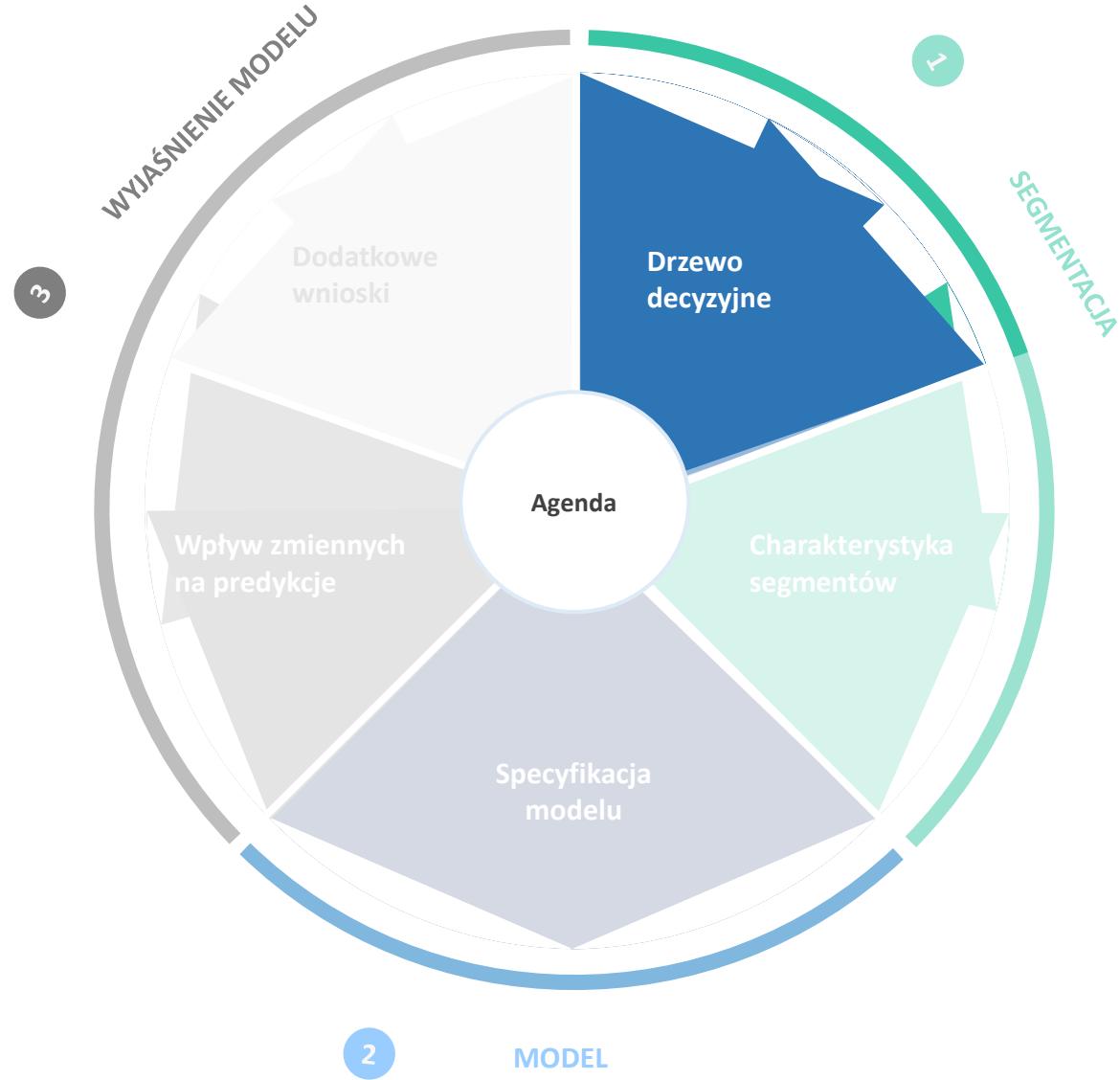
Piotr Halama

Robert Benke

Katarzyna Mucha

Mateusz Klimczak

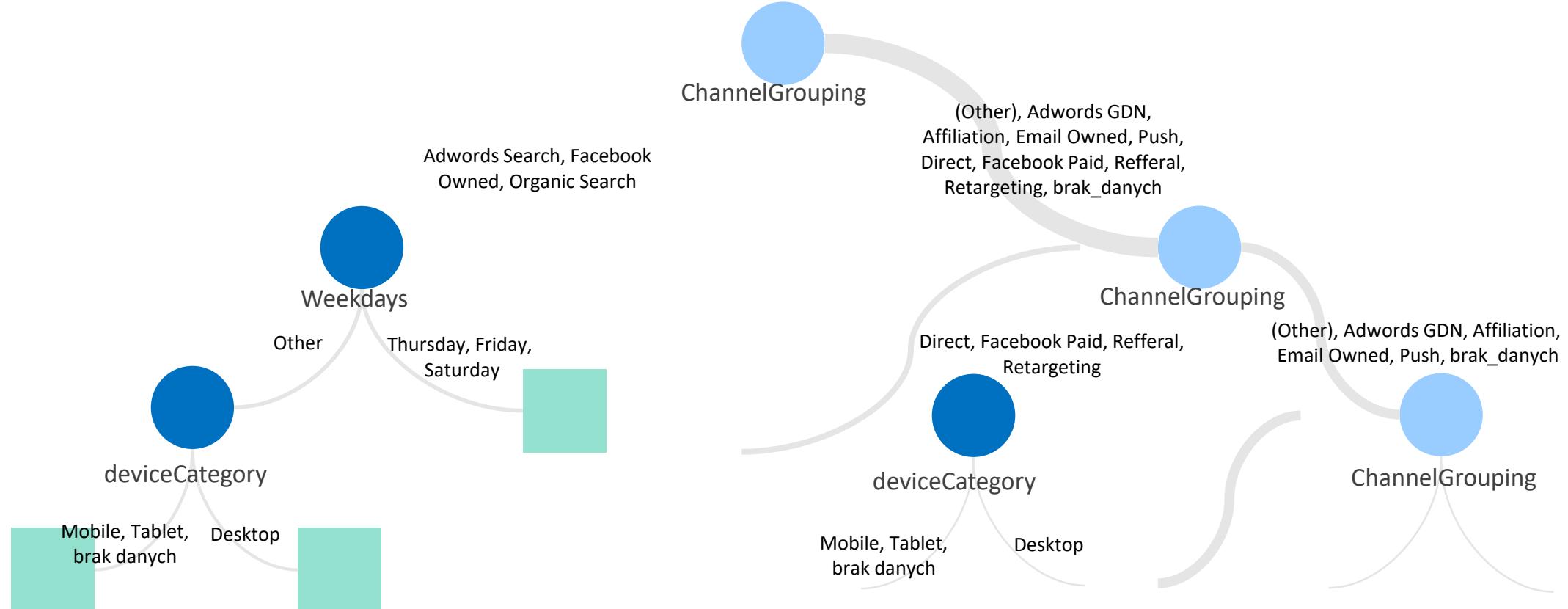
Tomasz Mikołajczyk



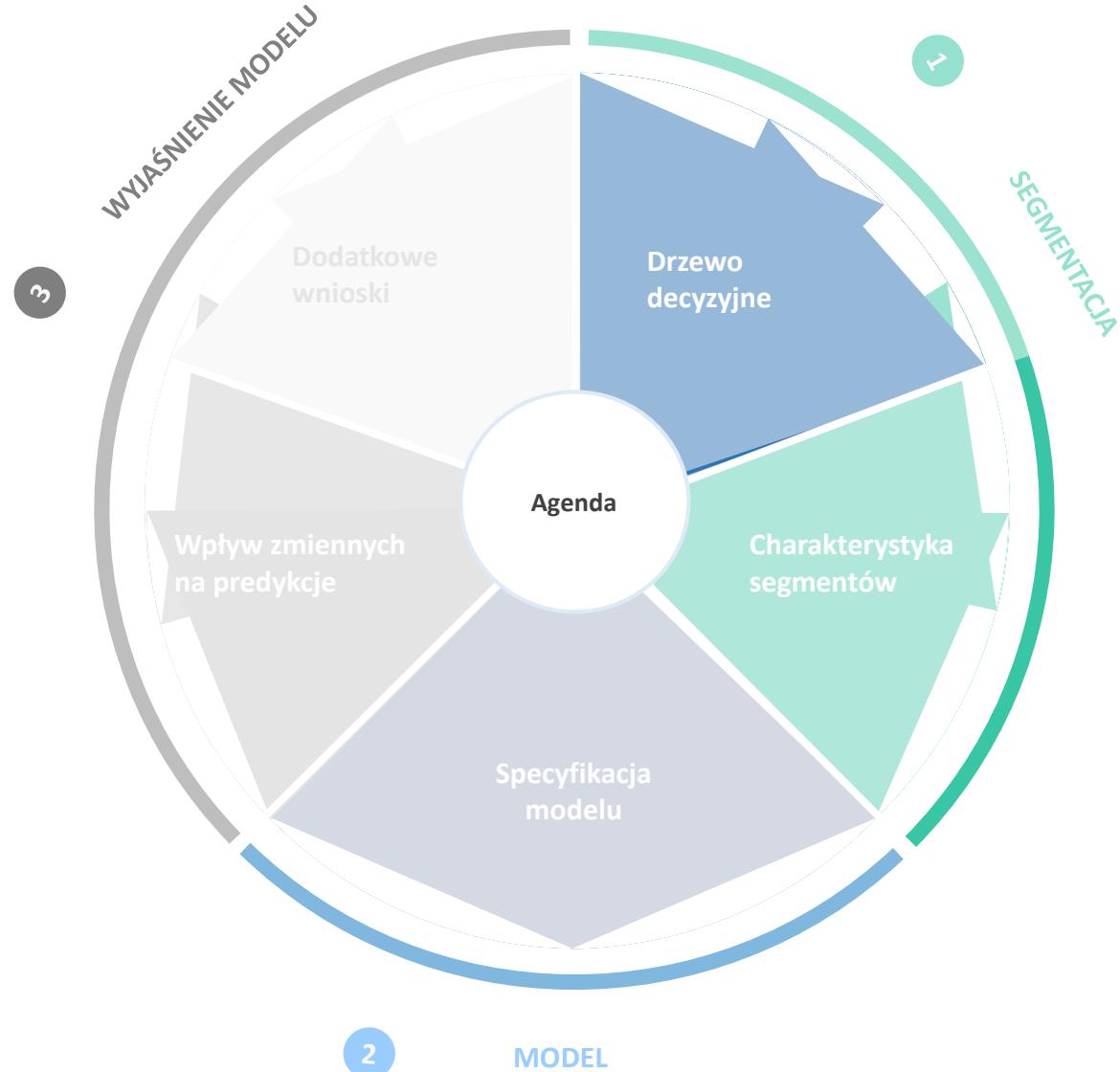
Przy użyciu drzewa decyzyjnego zostały wyznaczone segmenty klientów w pełni pokrywające użytkowników korzystających z portal LaModa

## Drzewo decyzyjne

Węzły      Węzły końcowe      Gałęzie



Segmenty zostały stworzone w oparciu o analizę drzewa decyzyjnego z uwzględnieniem potrzeb LaModa

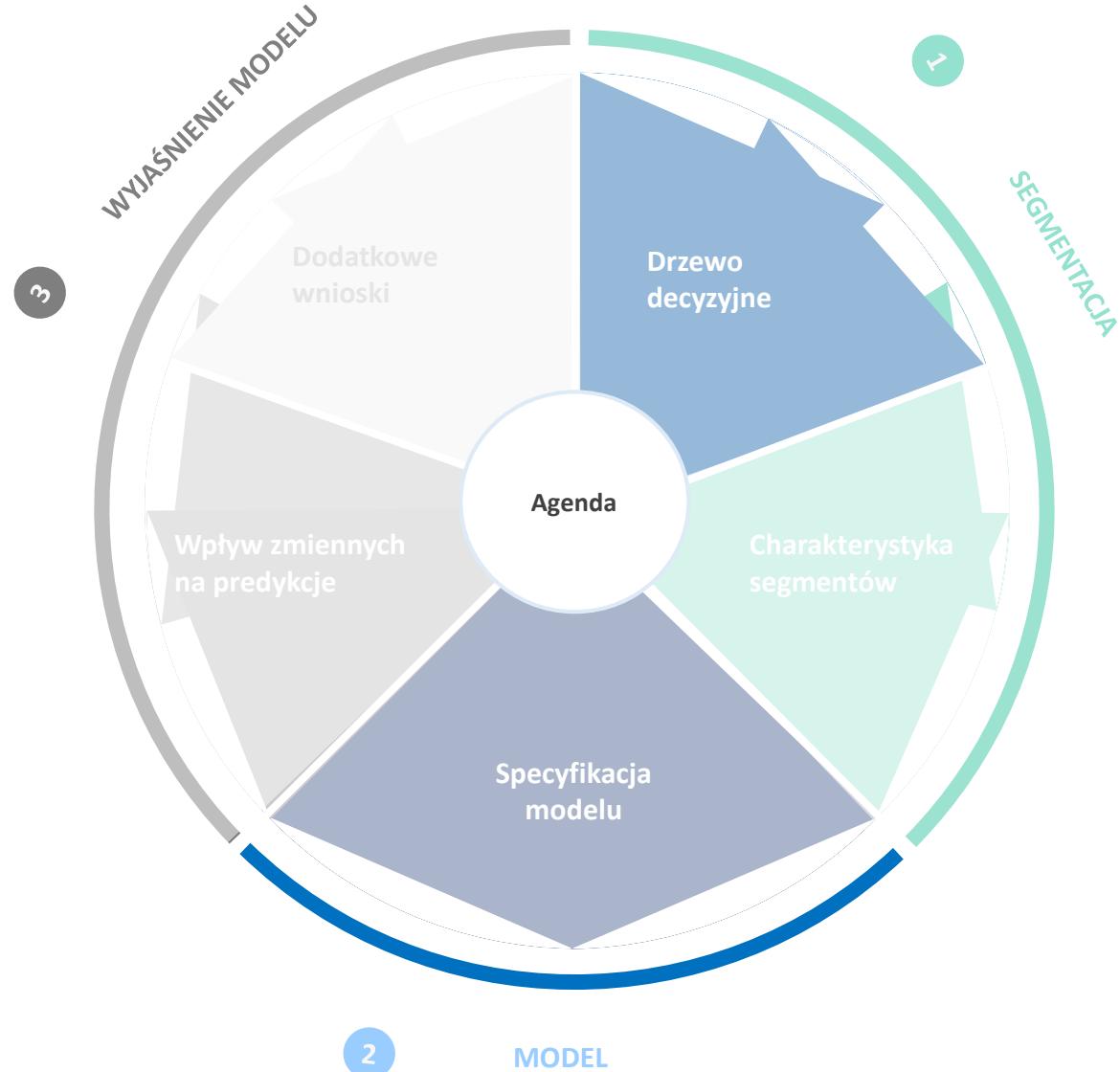


Po analizie wyników uzyskanych z podziału drzewa decyzyjnego zostały wybrane poniższe segmenty użyte w budowie modelu

## Charakterystyka segmentów

  Uzupełniające się agregaty

Segment	channelGrouping	Dzień tygodnia	deviceCategory	Udział w danych
Ograniczony kanałem	(Other), Adwords GDN, Affiliation, Email Owned, Push' +brak_danych	Mon-Sun	Wszystkie dostępne	30%
Weekend	Adwords Search, Facebook Owned, Organic Search	Thursday, Friday, Saturday	Wszystkie dostępne	25%
Dni pracujące	Adwords Search, Facebook Owned, Organic Search	Monday, Tuesday, Wednesday, Sunday	Wszystkie dostępne	30%
Użytkownik mobilny	Direct, Facebook Paid, Refferal, Retargeting	Mon-Sun	Mobile, Tablet + brak_danych	3%
Użytkownik stacjonarny	Direct, Facebook Paid, Refferal, Retargeting	Mon-Sun	Desktop	12%



Model został stworzony w oparciu o regresję liniową, aby możliwie uprościć proces jego implementacji w języku bazodanowym

## Budowa modelu

Model regresji liniowej cechuje się wysoką stabilnością, jednak słabo dopasowuje się do bardzo złożonych zależności

Dzięki poniższym rozwiązaniom cecha ta została usunięta

Interakcje zmiennych  
niezależnych  
np.  $y=x_1*x_2$

Iloczyny dwóch, trzech, wielu  
zmiennych

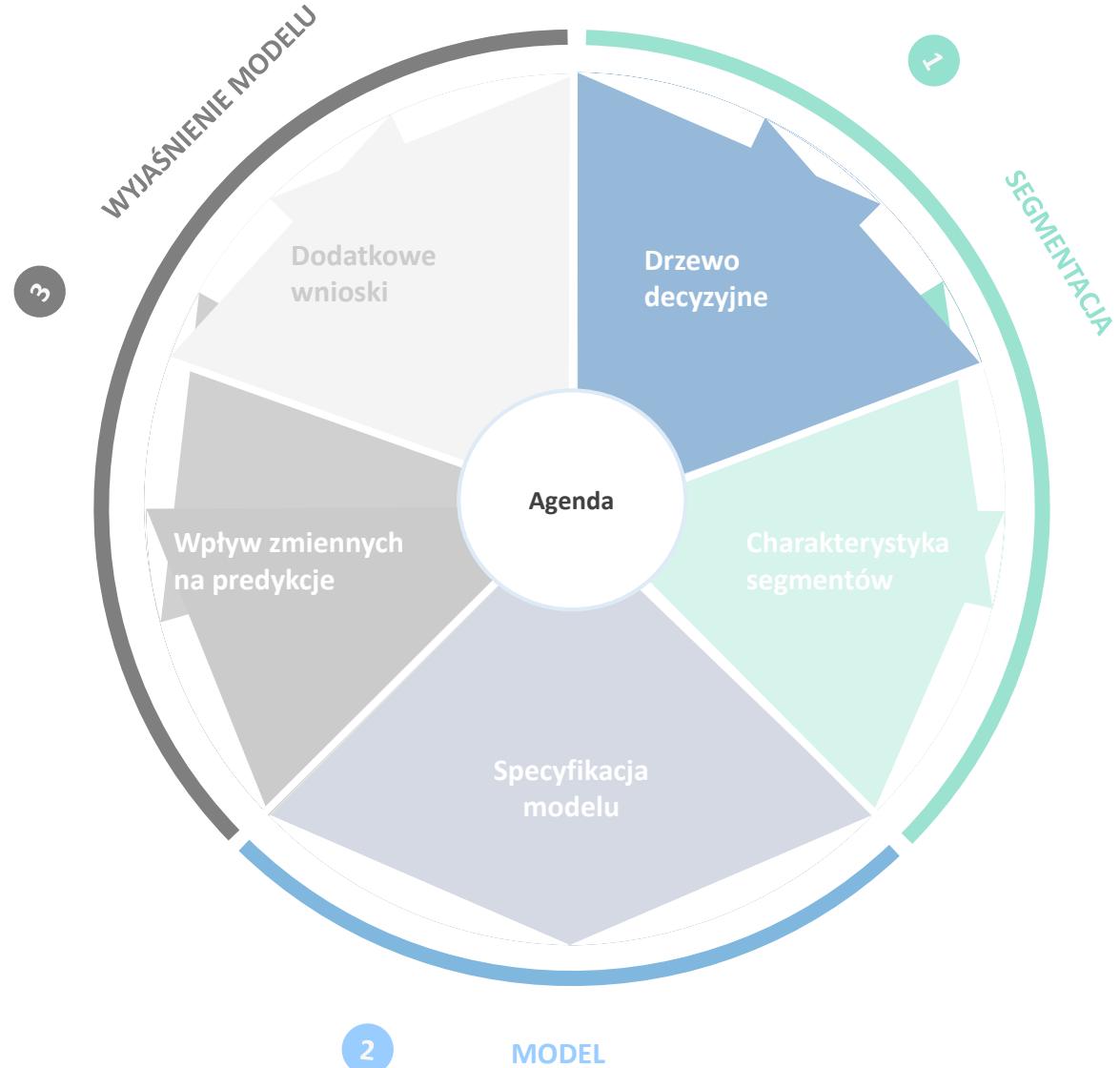
Kwadraty i sześciany zmiennych

### Zmienne wykorzystane w modelu

	subcategory_id_mapped		transactions
	brand_mapped		price
	discount_percent		category_id
	base_price		

Zmienne do modelu zostały wybrane na podstawie testu t-studenta przy poziomie istotności 5%. Zmienne użyte w modelu nie były współliniowe

Przy dużej ilości segmentów model spełniał założenia dotyczące rozkładu reszt oraz braku współliniowości zmiennych. Zmniejszenie ilości segmentów skutkowało odmiенноściami reszt od rozkładu normalnego

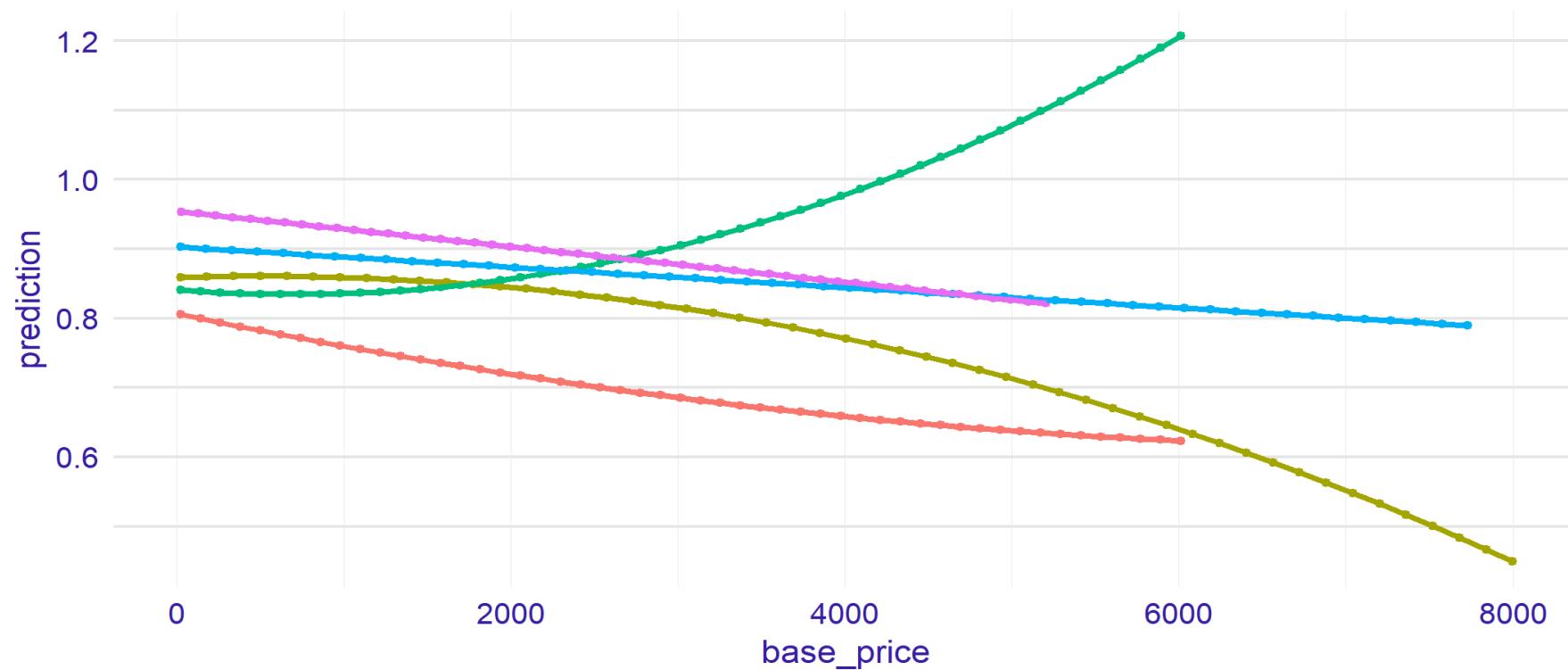


Analiza wszystkich segmentów

Większość segmentów negatywnie reagowała na zmianę ceny produktu

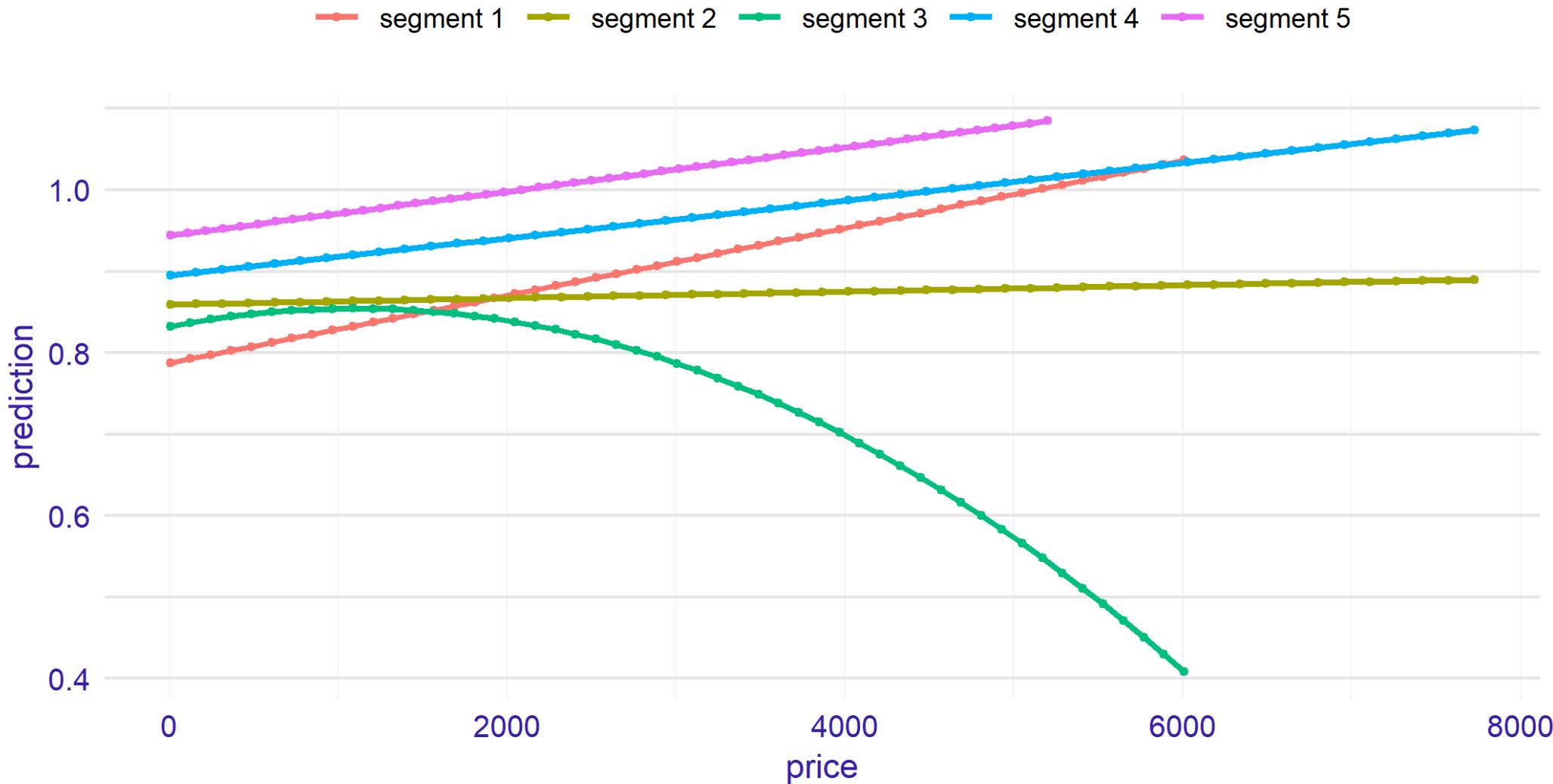
PD – plot

segment 1 segment 2 segment 3 segment 4 segment 5



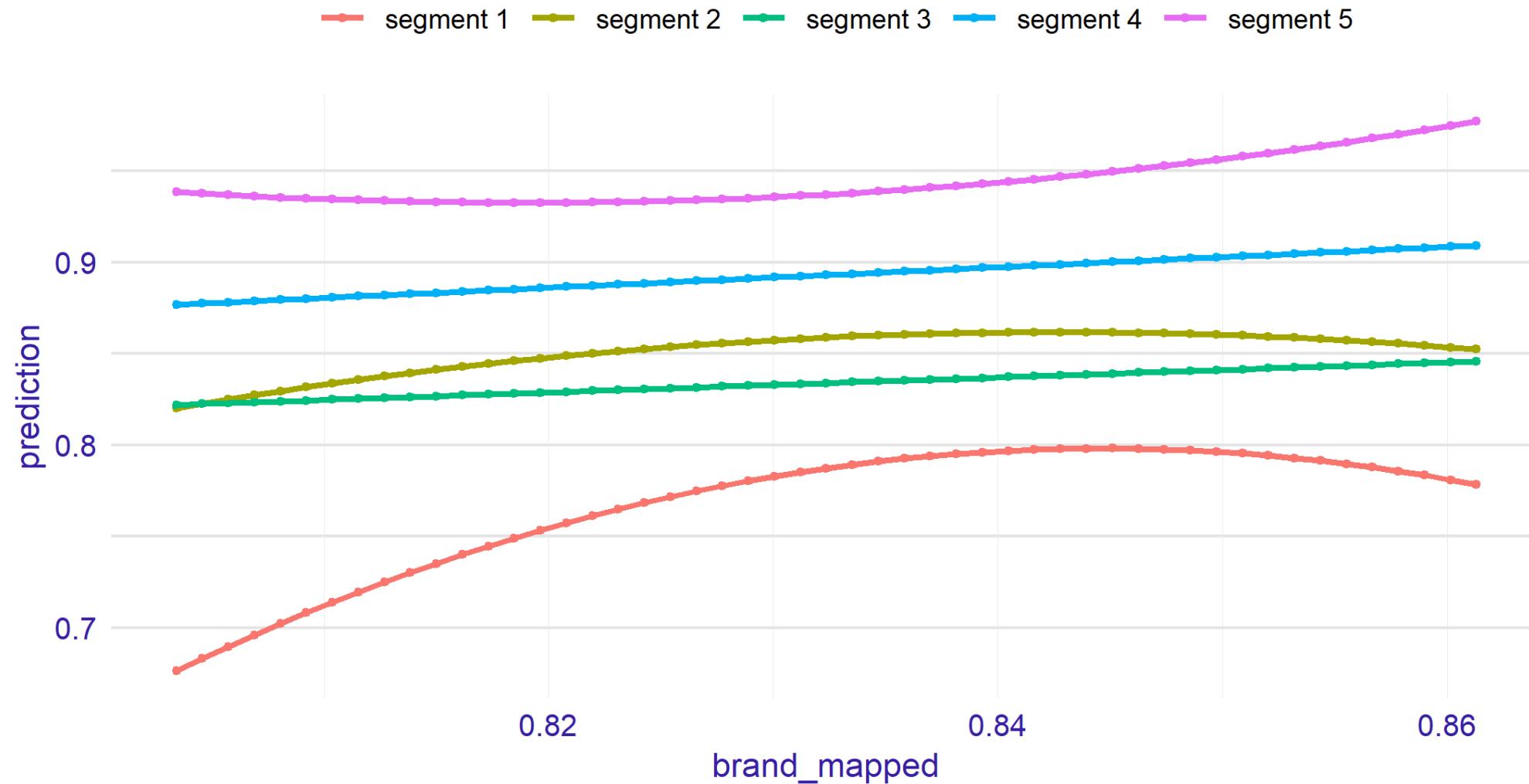
Klienci z segmentu 3ego reagują odwrotnie niż pozostali. Z kolei segment drugi najsilniej reaguje na zwiększenie ceny

## PD - plot for models for all segments



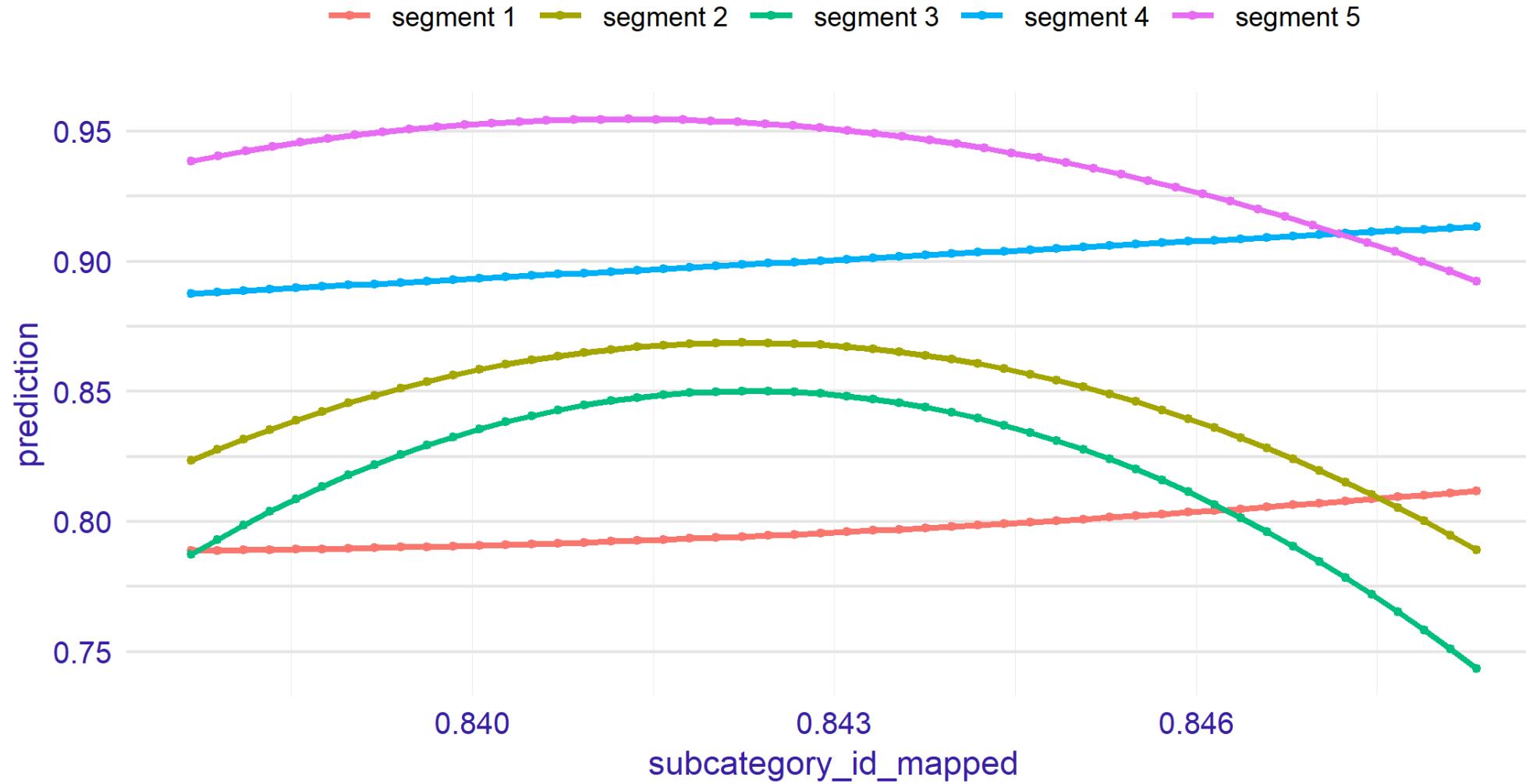
Z wykresów PDP dla price i base\_price możemy wywnioskować niechęć konsumentów do kupowania produktów drogich chyba że produkty te są wysoko przecenione.

# PD - plot for models for all segments



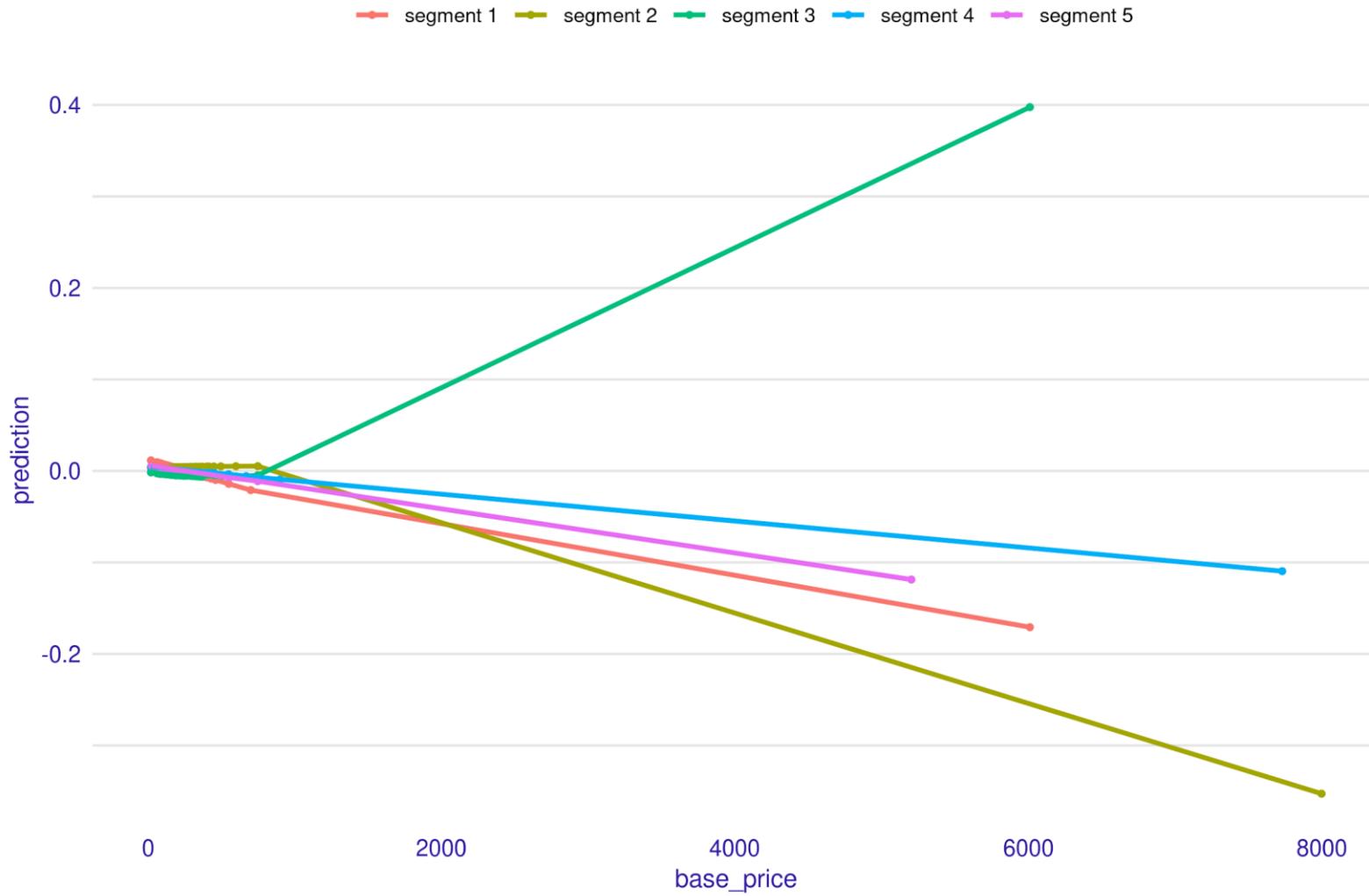
Wykres PD dla zmiennej `brand_mapped` (trend rosnący) ma spodziewany kształt dla taget encoding tzn. kategorie z wyższą średnią klikanością dostają też wyższy score.

## PD - plot for models for all segments



Zastanawiająca jest natomiast kwadratowa charakterystyka PDP dla zmapowanej zmiennej category. Intuicyjnie zależność ta powinna mieć charakter liniowy, dodatkowo implikacją postaci kwadratowej jest bardzo niski score dla kategorii o bardzo wysokim współczynniku konwersji.

## ALE - plot for models for all segments



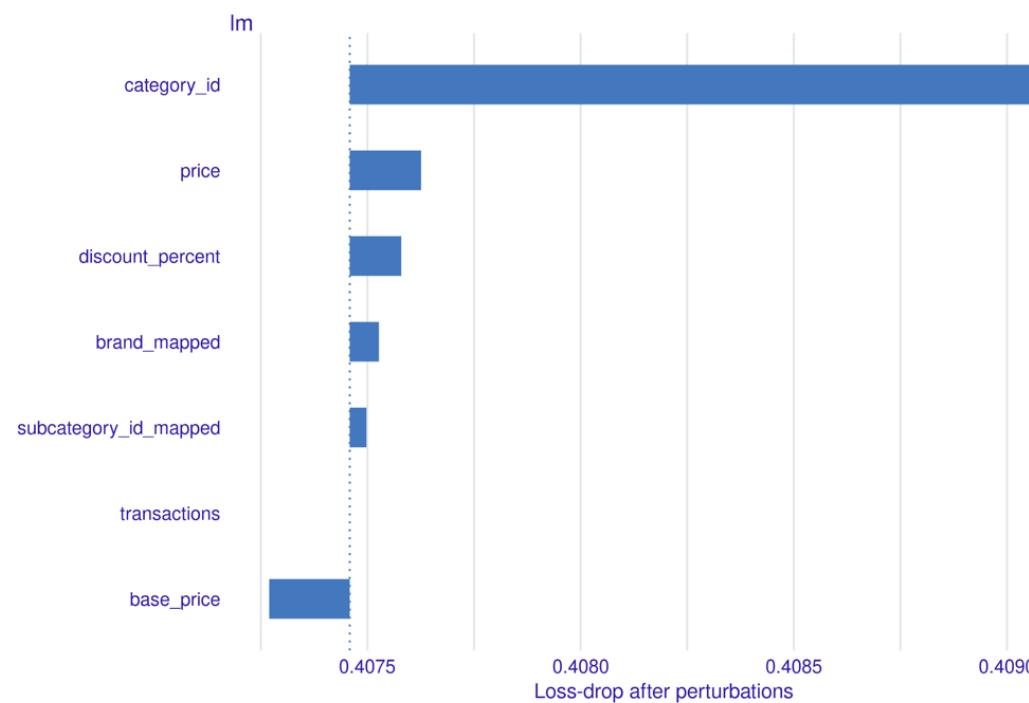
ALE plot, stworzony po kwantylach wartosci nie wnosi żadnej informacji, wręcz przeciwnie, nie widać na nim kwadratowej zależności zmiennej objaśniającej od zmiennej celu.

Analiza per segment  
Segment 1

Dzięki analizie poniższych wykresów można stwierdzić, że zmienna kategoryczna ma największy wpływ na predykcje modelu dla segmentu klientów ograniczonych kanałem wejścia na stronę

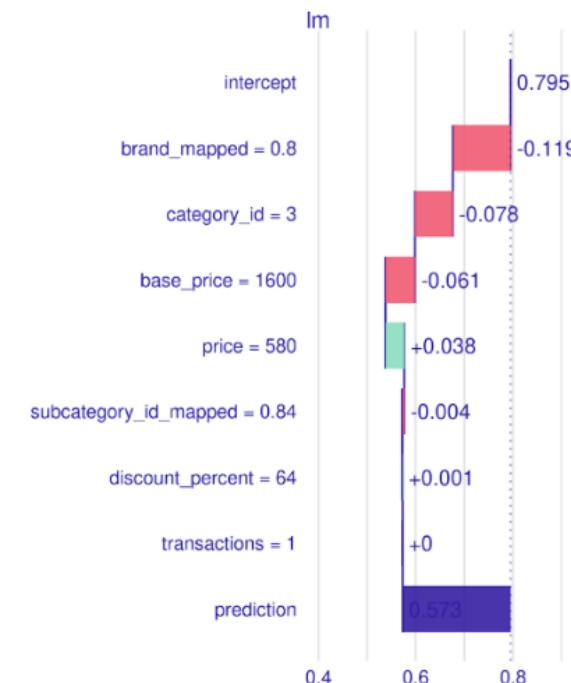
## Segment klientów ograniczonych kanałem wejścia na stronę

### Istotność poszczególnych zmiennych

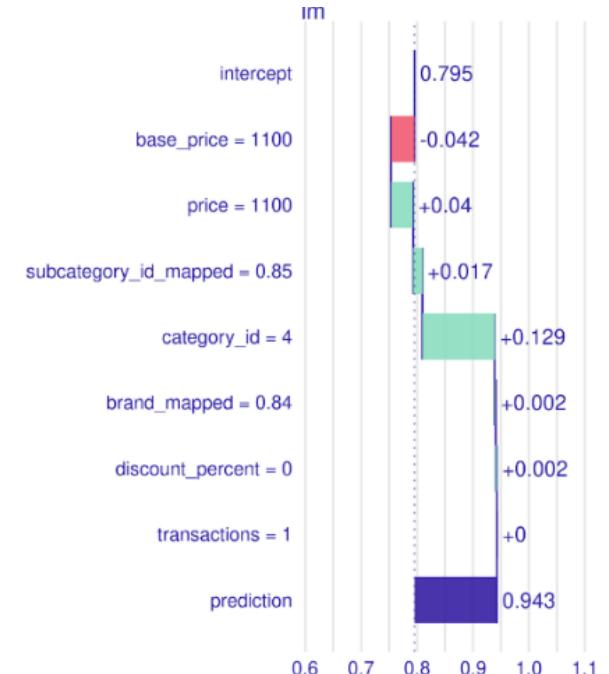


### Wahania predykcji modelu

#### Przykład najniższego wyniku



#### Przykład najwyższego wyniku

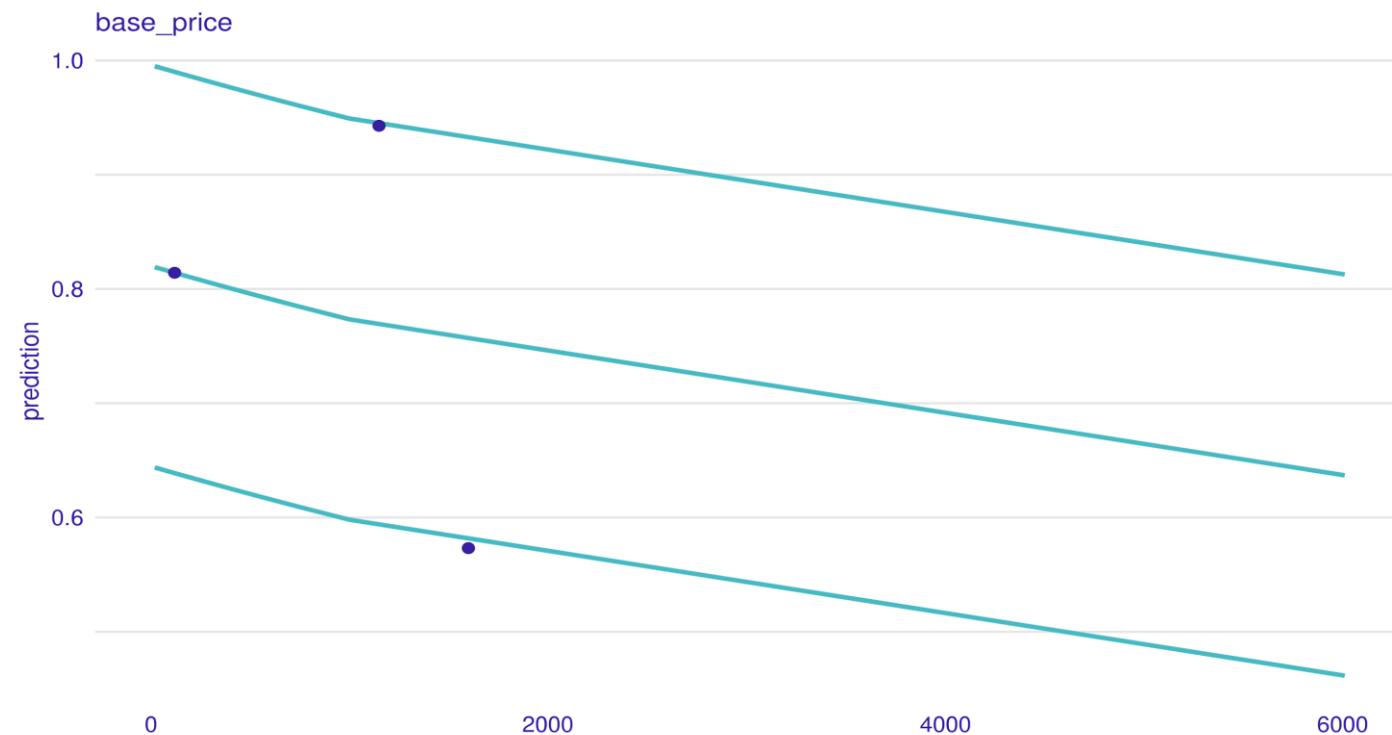


Wyniki predykcji potwierdzają, że najistotniejsza jest zmienna kategoryczna

Im droższe produkty, tym klienci z segmentu ograniczonego kanałem wejścia na stronę, będą nimi mniej zainteresowani

### Ceteris Paribus | Segment klientów ograniczonych kanałem wejścia na stronę

Najwyższy wynik/ losowy/ Najniższy wynik

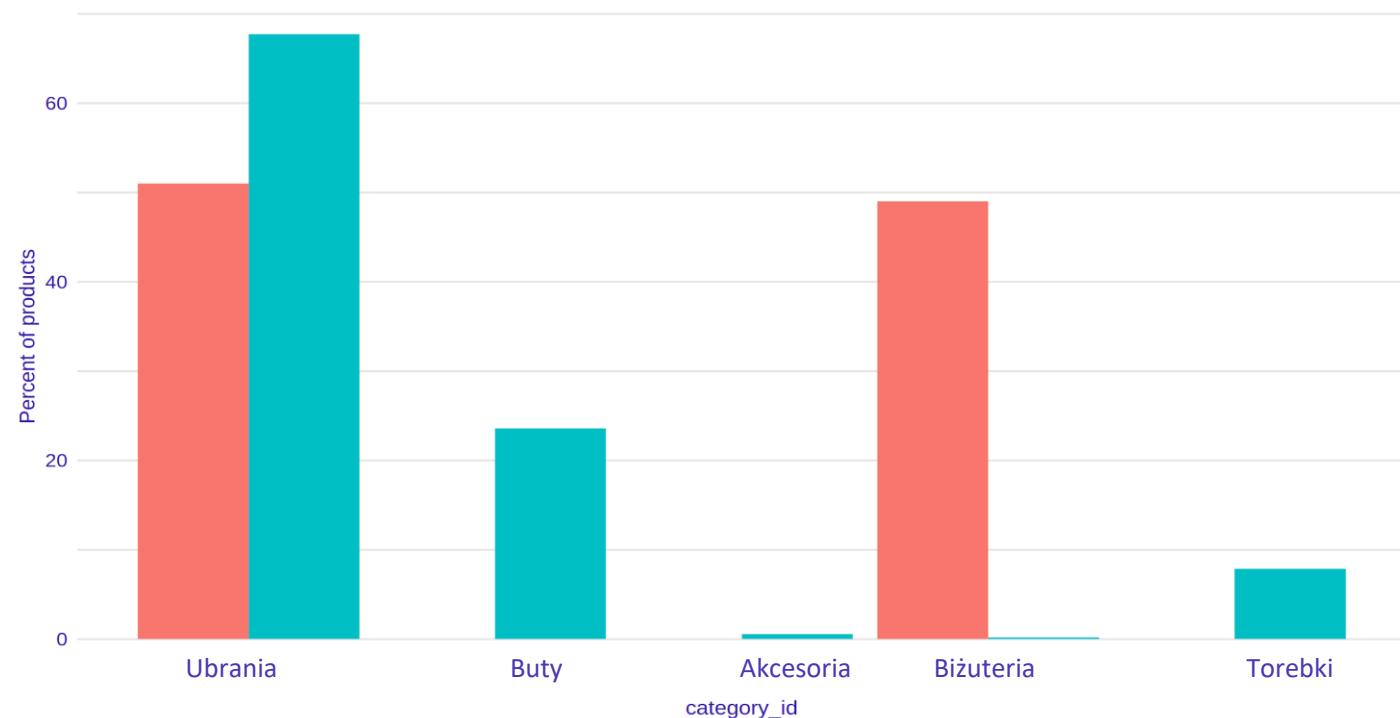


- Przy innych czynnikach niezmienionych, zmiana ceny produktu powoduje spadek wyniku predykcji o około 0,2.
- Charakter (liniowy) odpowiedzi modelu przy zmianie ceny nie różni się w zależności od pozycji jaką dostał produkt. Czynniki wpływające na pozycjonowanie nie posiadają silnych interakcji z ceną.

Klienci segmentu ograniczeni kanałem wejścia na stronę są najbardziej zainteresowani ubraniem butami oraz torebkami, jednak najczęściej klikają w ubrania oraz biżuterię

### Porównanie category\_id | Segment klientów ograniczonych kanałem wejścia na stronę

Top 100    Wszystkie

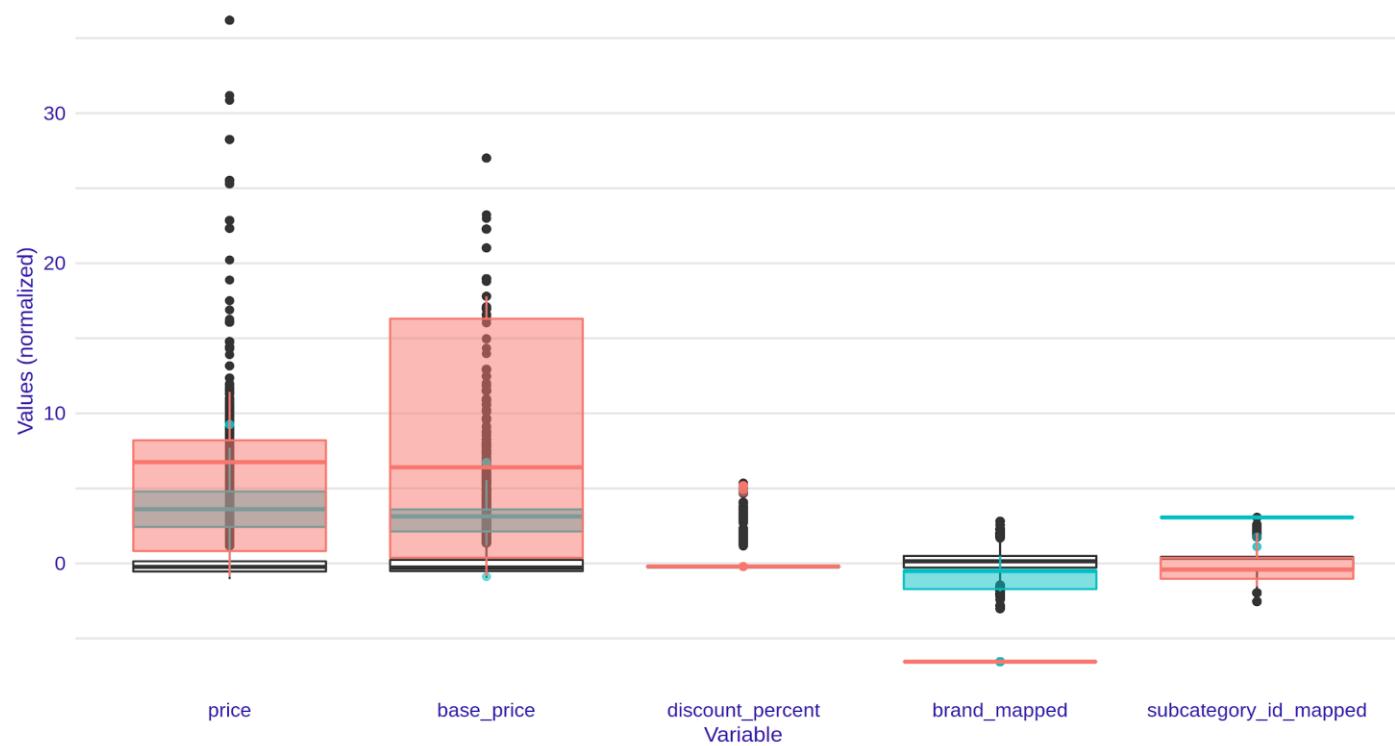


- Najczęściej wyświetlane są ubrania i buty, a porównywalnie często jak produkty z kategorii ubrania kupowana jest biżuteria.
- Pierwszy segment to niemal wyłącznie ubrania i biżuteria. Segment drugi zawiera przede wszystkim biżuterię, a pozostałe trzy są bardziej zrównoważone z przewagą butów i torebek.

Poniższy wykres przedstawia porównanie 50 najwyższych oraz 50 najniższych obserwacji z rozkładem wszystkich obserwacji

### Response box plot | Segment klientów ograniczonych kanałem wejścia na stronę

Bottom 50 Top 50

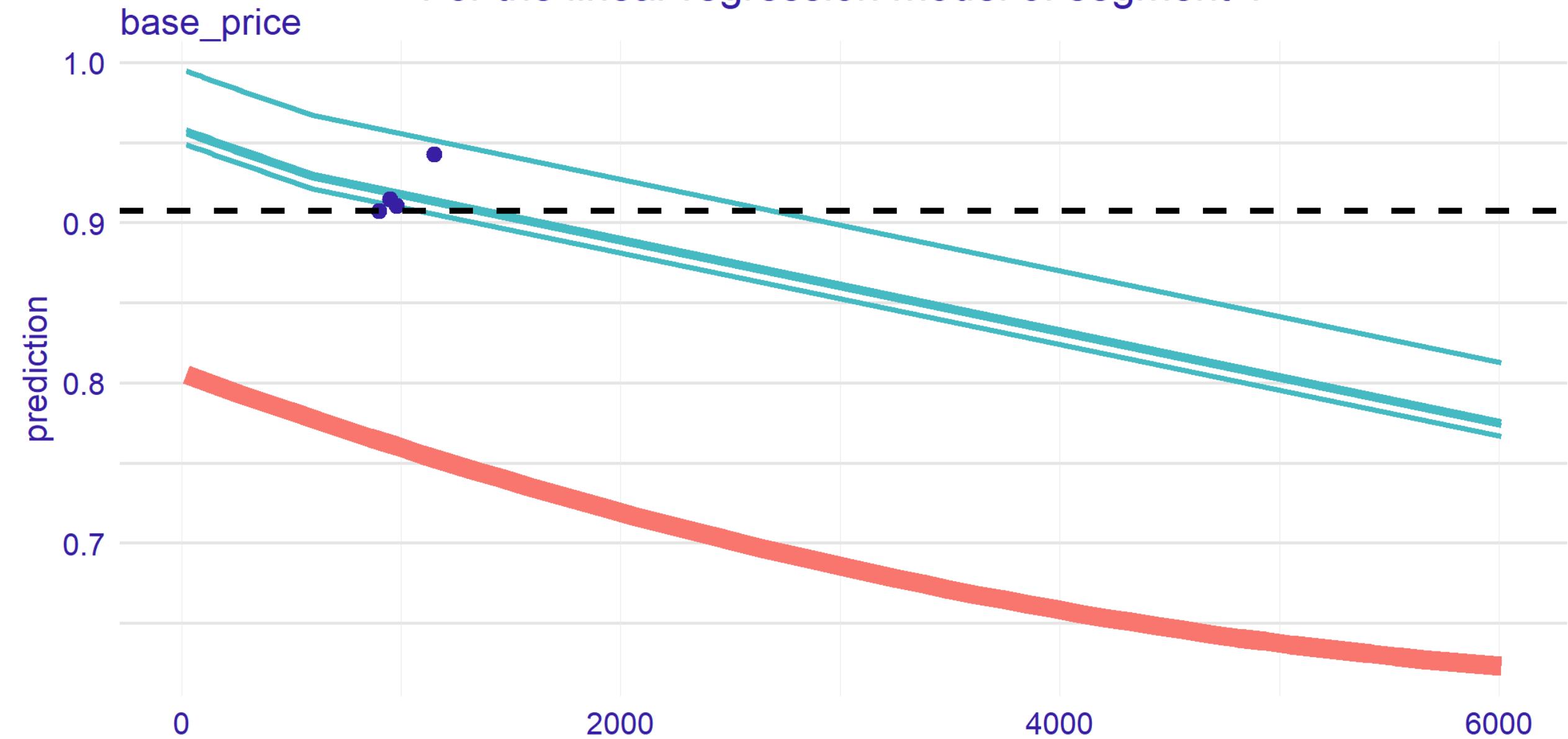


Klienci z segmentu ograniczonego kanałem wejścia na stronę przeważnie nie reagują na zniżki

- Na następnych slajdach pokazane są wykresy Ceteris Paribus dla dziesięciu najwyżej ocenionych przez model produktów w porównaniu do wykresu PDP dla wszystkich rekordów.
- Widzimy również do jakiej wartości zmienna objaśniana musiała spaść aby produkt znalazł się poza pierwszą dziesiątką.
- Najszybciej pozycja produktów zmieniała się przy manipulowaniu ceną natomiast najwolniej w przypadku zmapowanej zmiennej kategorii produktu. Umiarkowany wpływ mają zmienne brand\_mapped i base\_price.

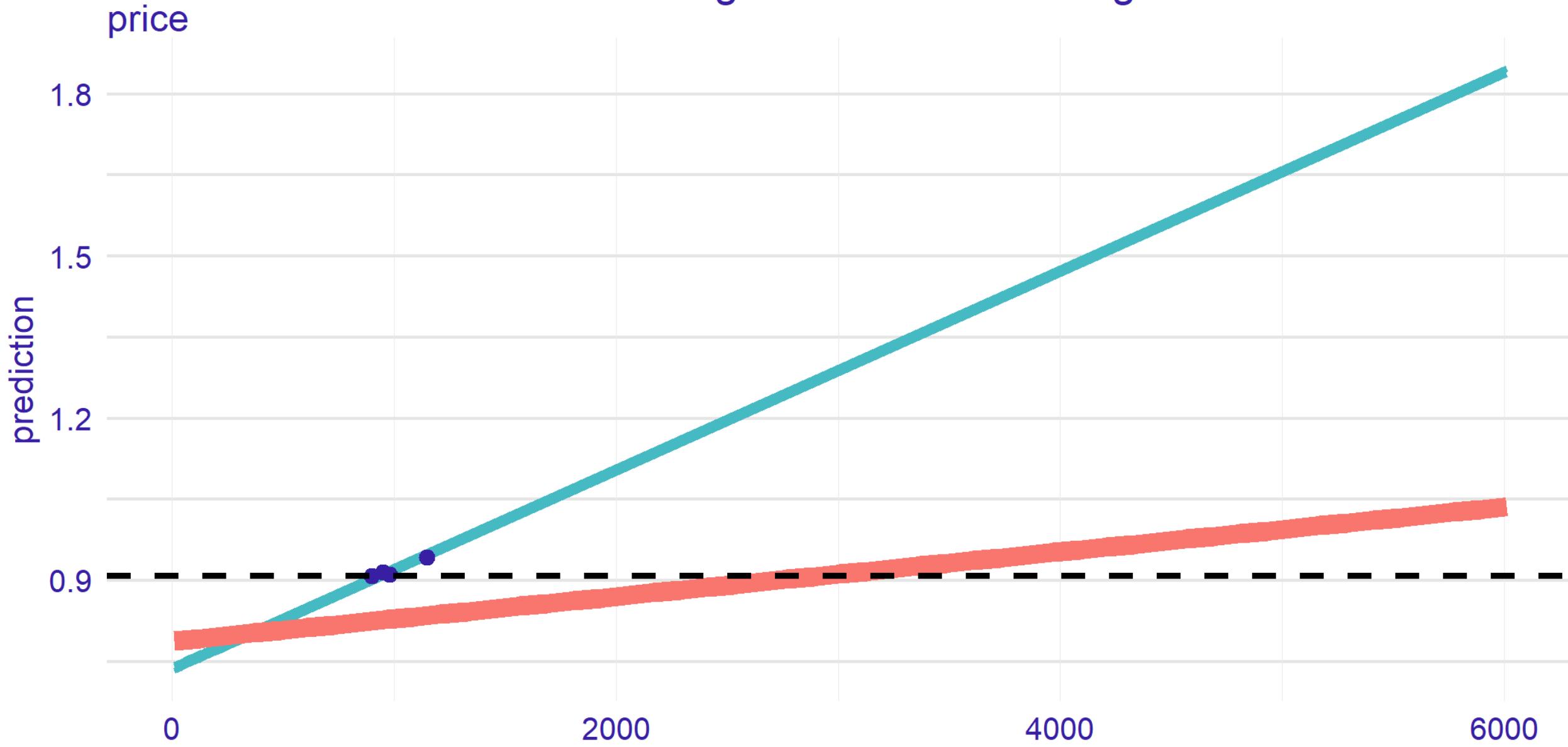
# Ceteris Paribus Profiles (top 10 responses) + PD - plot

## For the linear regression model of segment 1



# Ceteris Paribus Profiles (top 10 responses) + PD - plot

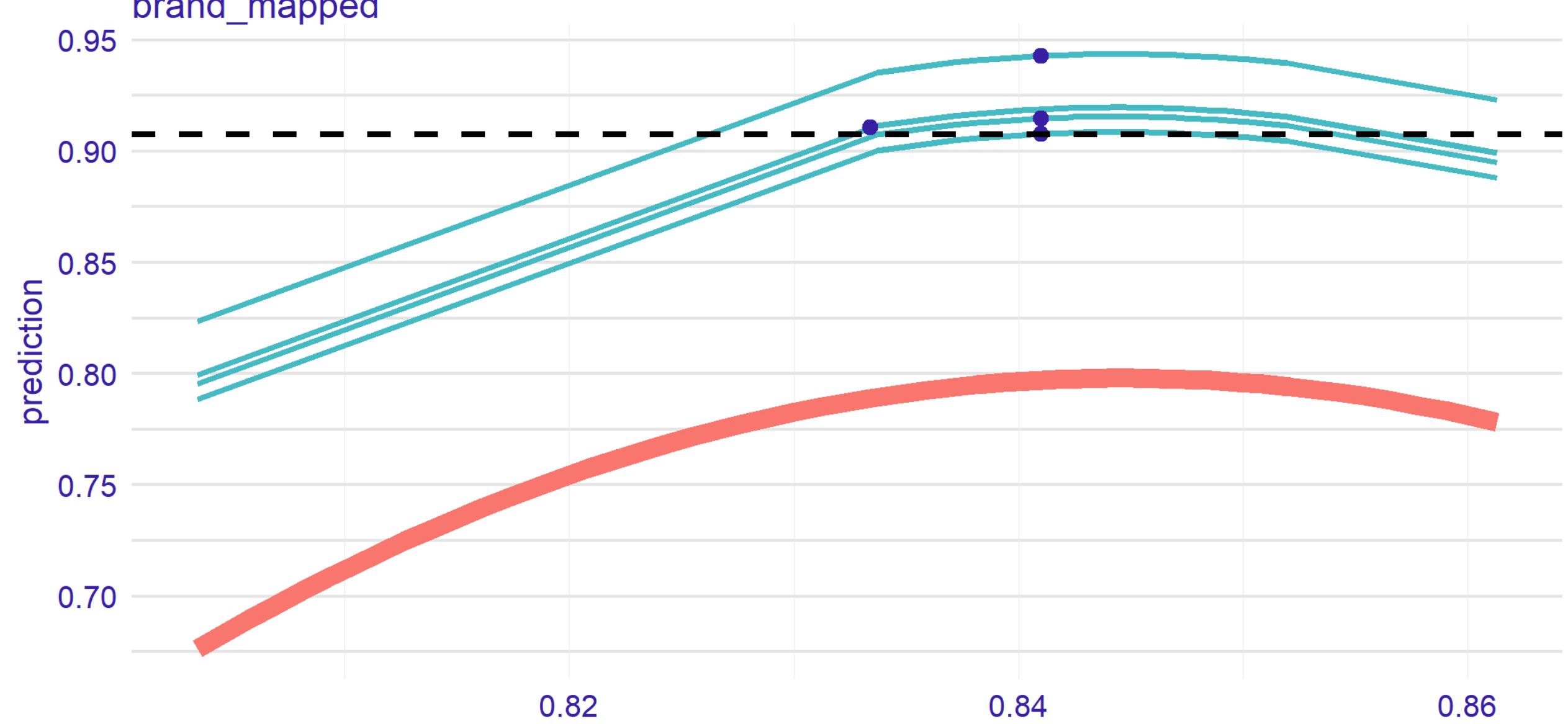
For the linear regression model of segment 1



# Ceteris Paribus Profiles (top 10 responses) + PD - plot

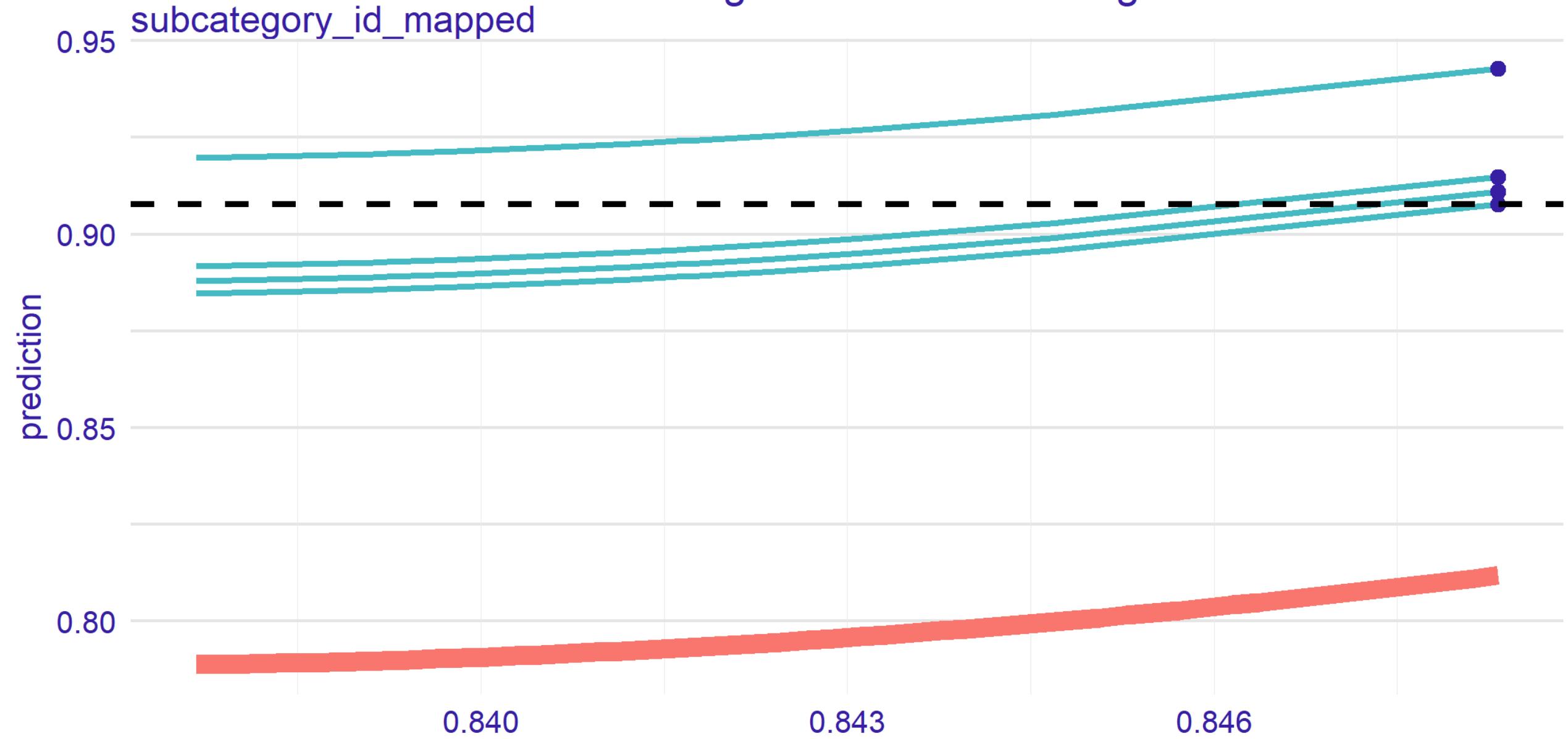
## For the linear regression model of segment 1

brand\_mapped



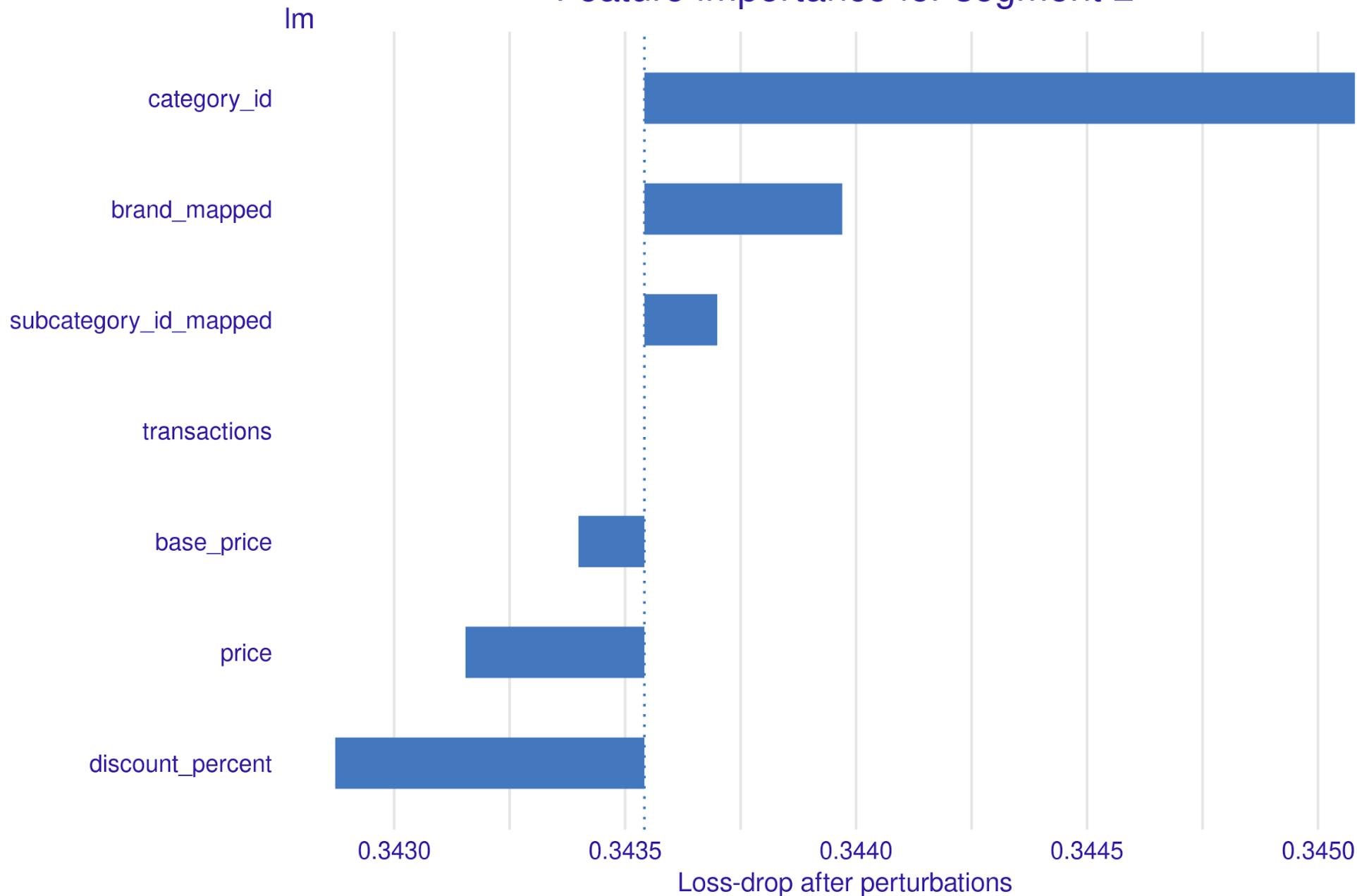
# Ceteris Paribus Profiles (top 10 responses) + PD - plot

## For the linear regression model of segment 1



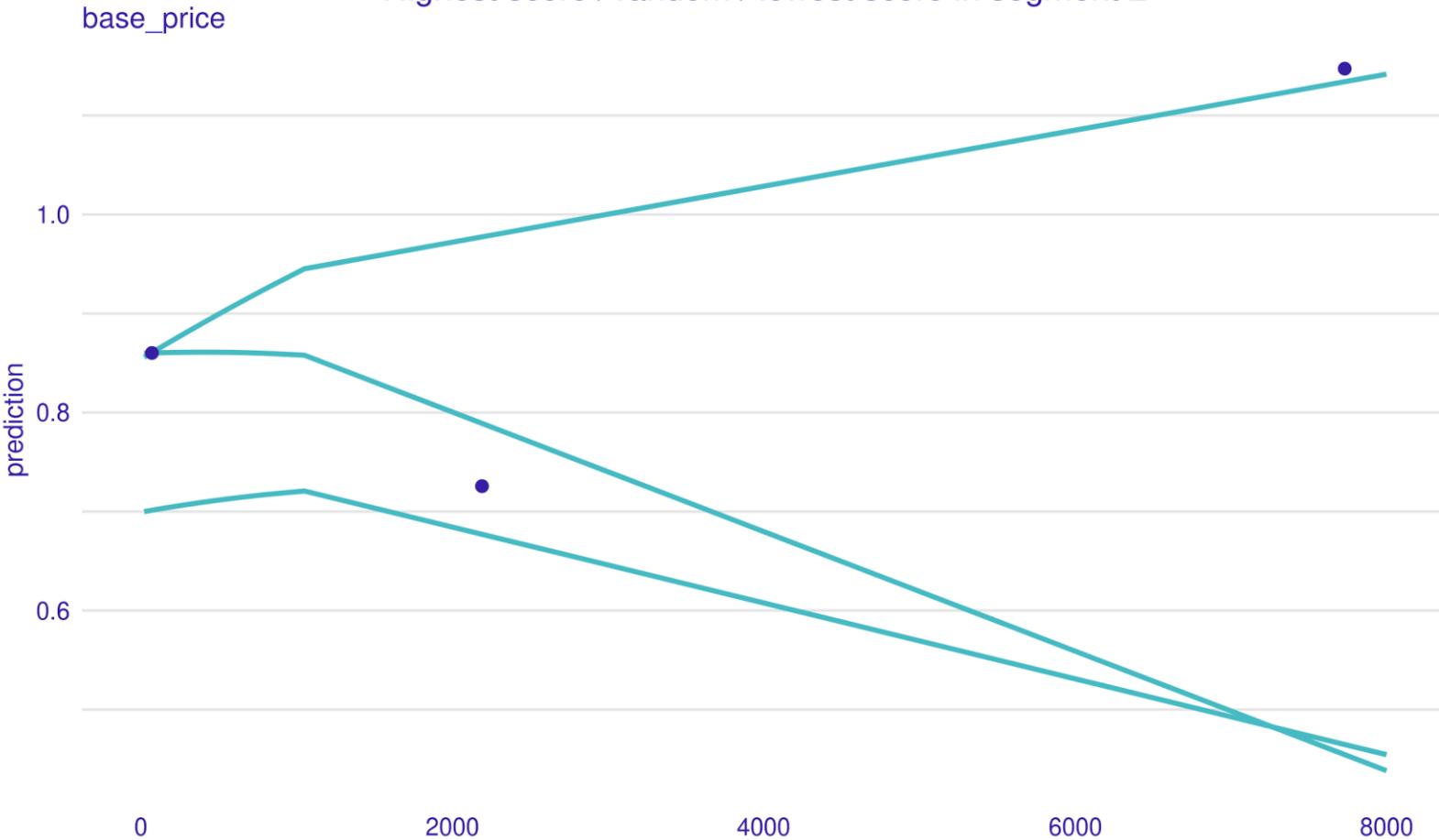
Analiza per segment  
Segment 2

## Feature importance for segment 2



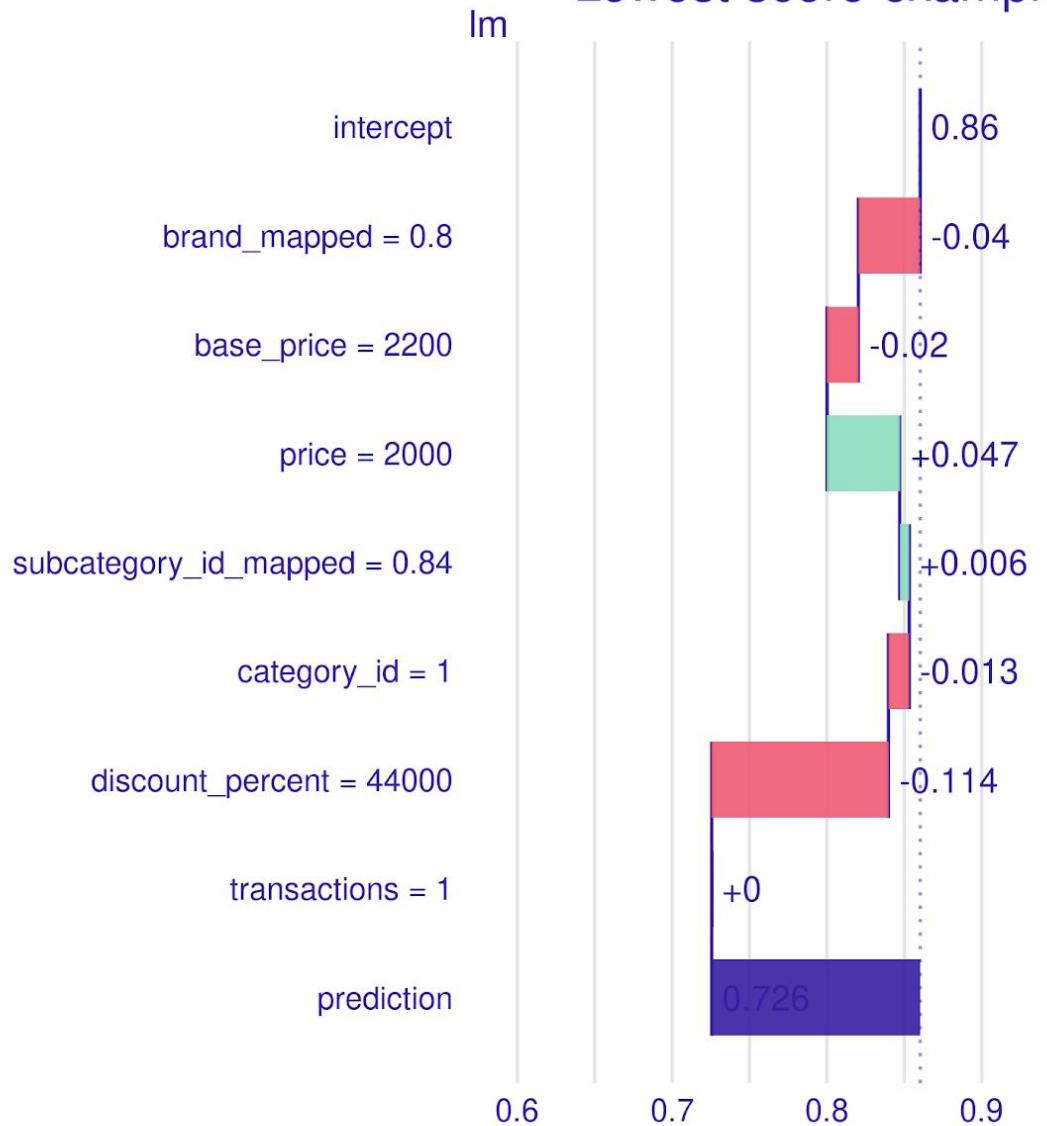
## Ceteris Paribus Profiles

Highest score / random / lowest score in segment 2

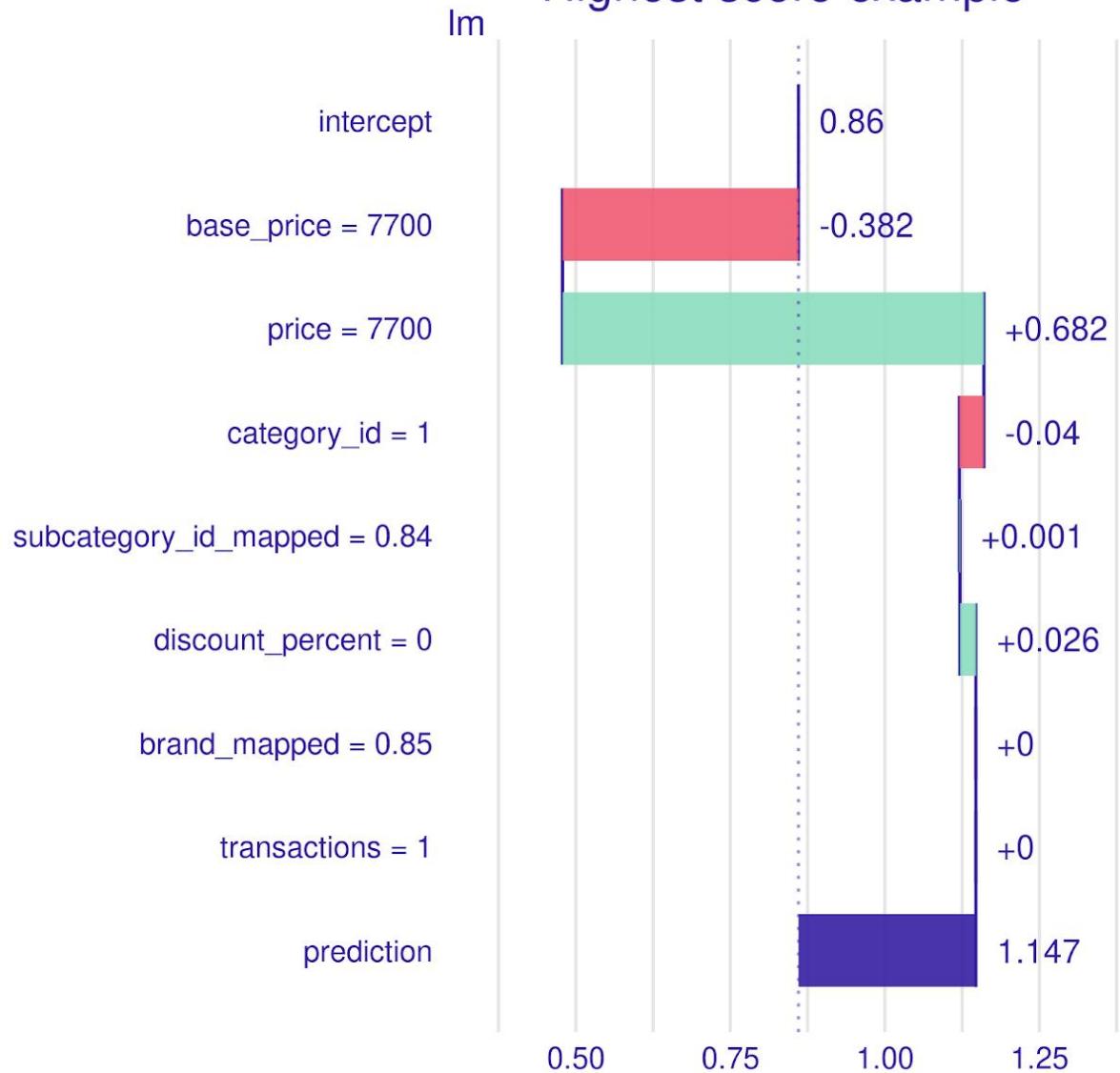


- Segment drugi to również klienci szukający względnie tanich produktów, raczej nie przecenianych.
- Odpowiedź modelu przy zmianie ceny mocno zależy od innych cech produktu (wzrost ceny może być odebrany zarówno pozytywnie jak i negatywnie).

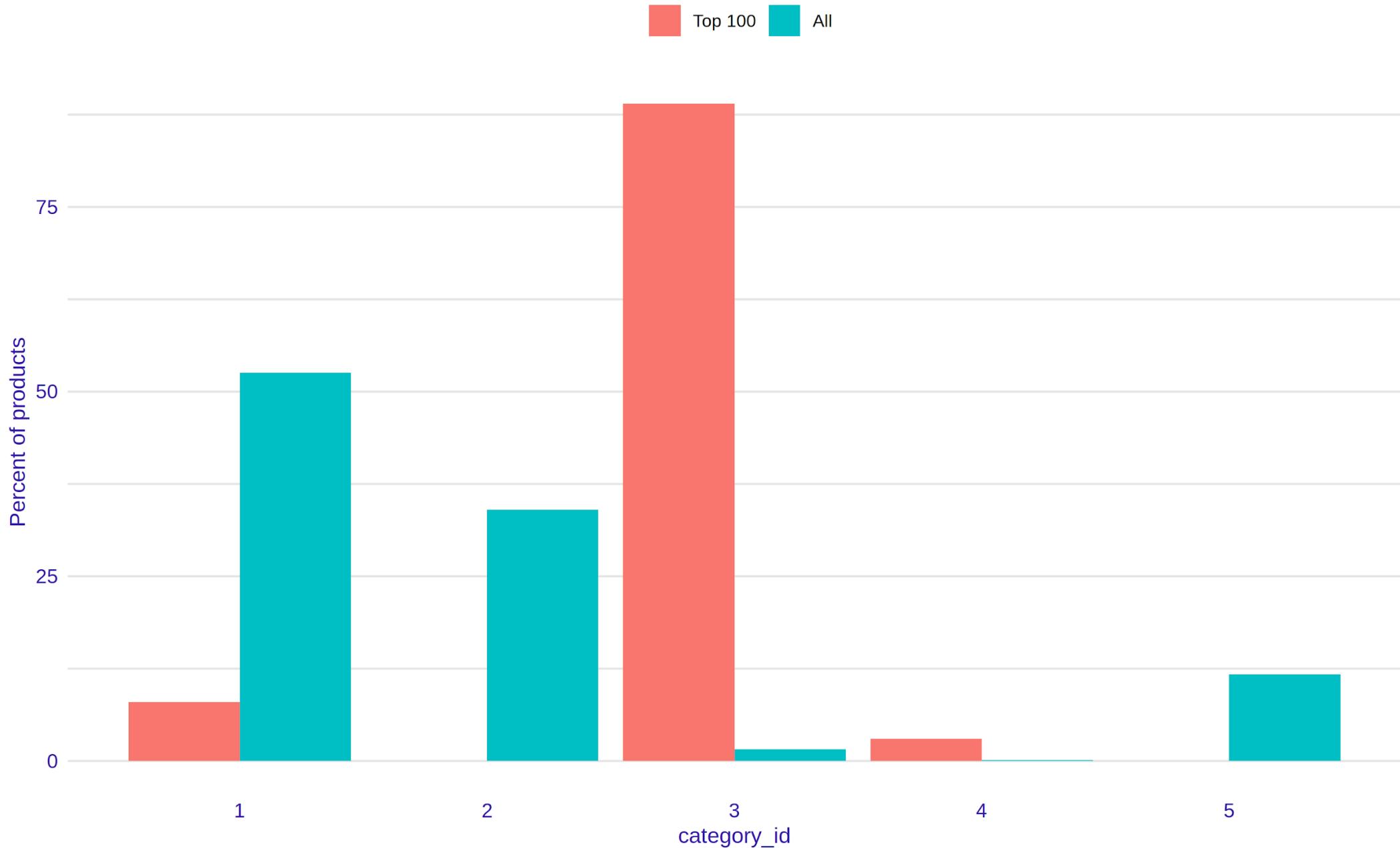
## Lowest score example



## Highest score example

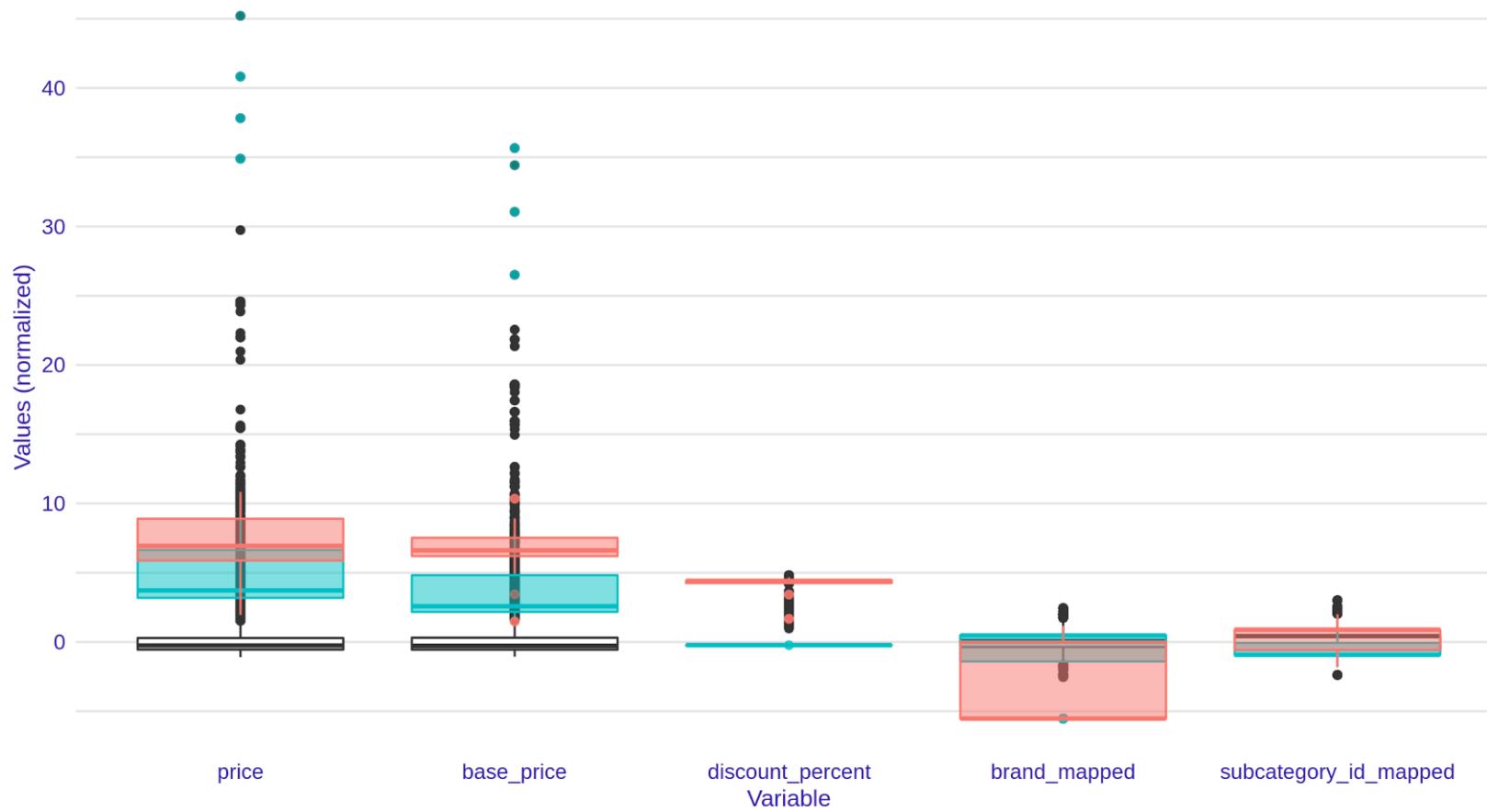


## Comparison of category\_id values in segment 2



## Response box plot for segment 2

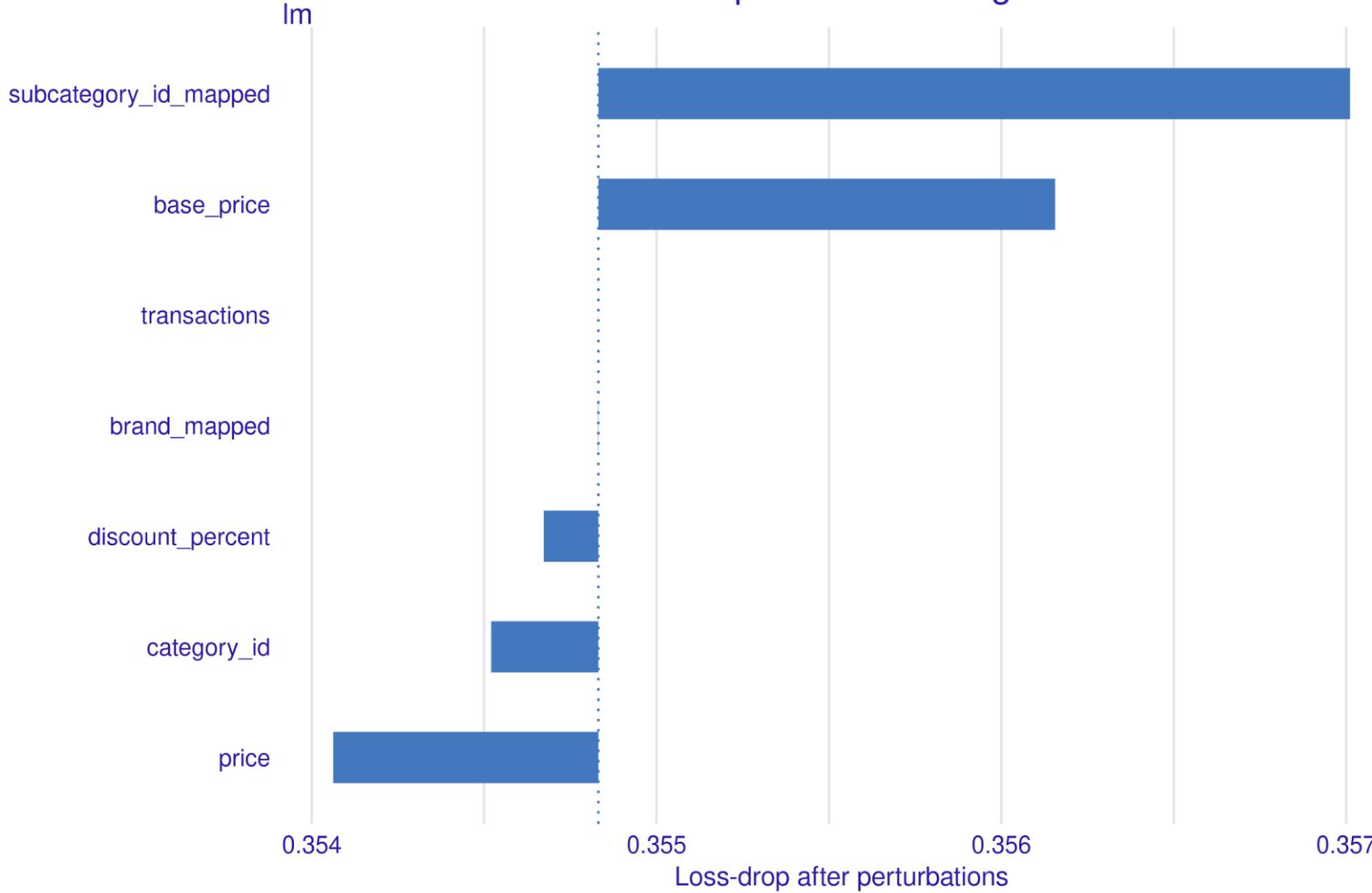
Bottom 50 Top 50



Responce box pokazuje jeszcze lepiej to co udało nam się zauważyc przy użyciu Ceteris Paribus i variable importance, mianowicie, segment 2 promuje produkty tańsze których ceny nie są obniżane.

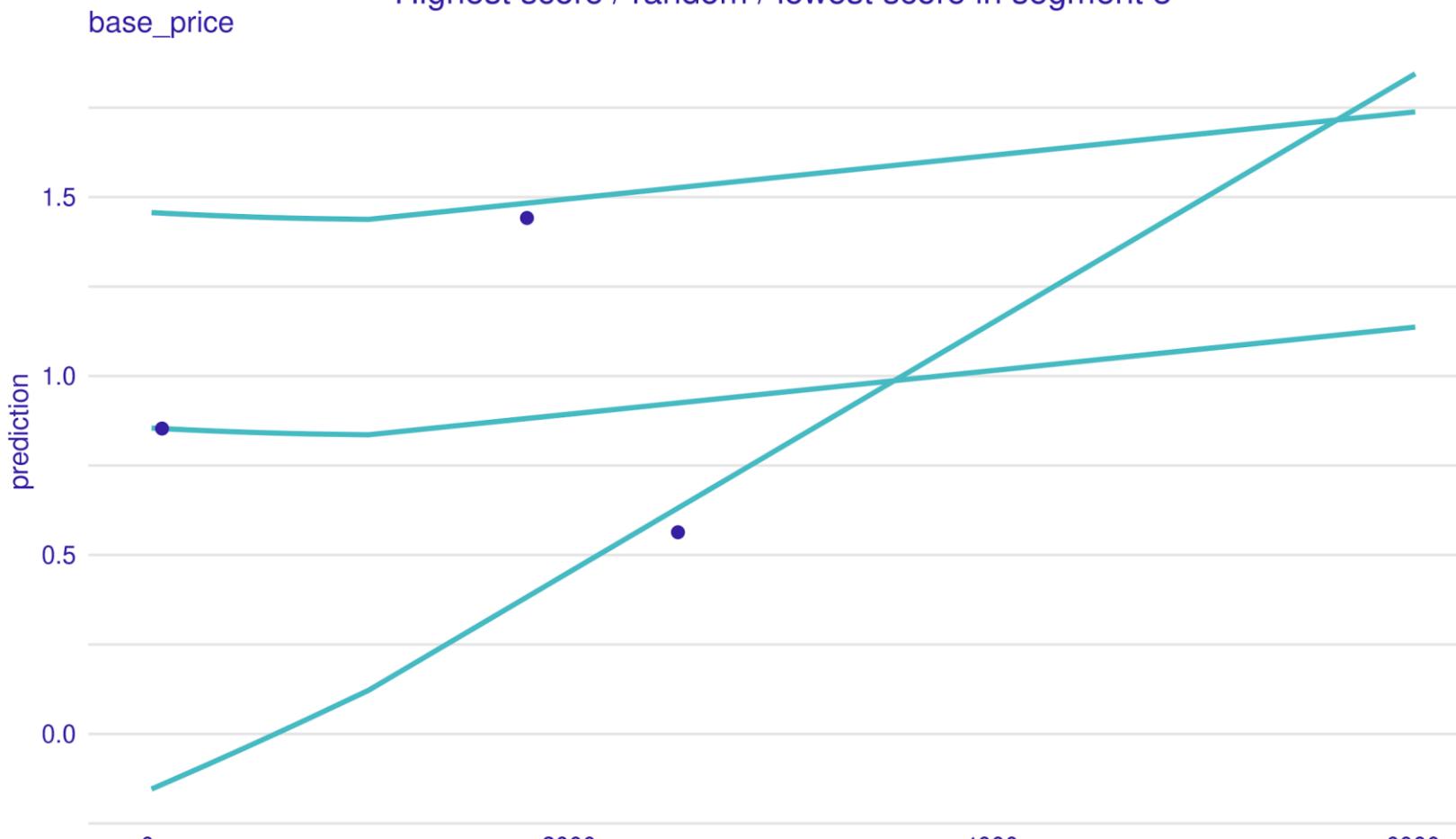
Analiza per segment  
Segment 3

### Feature importance for segment 3

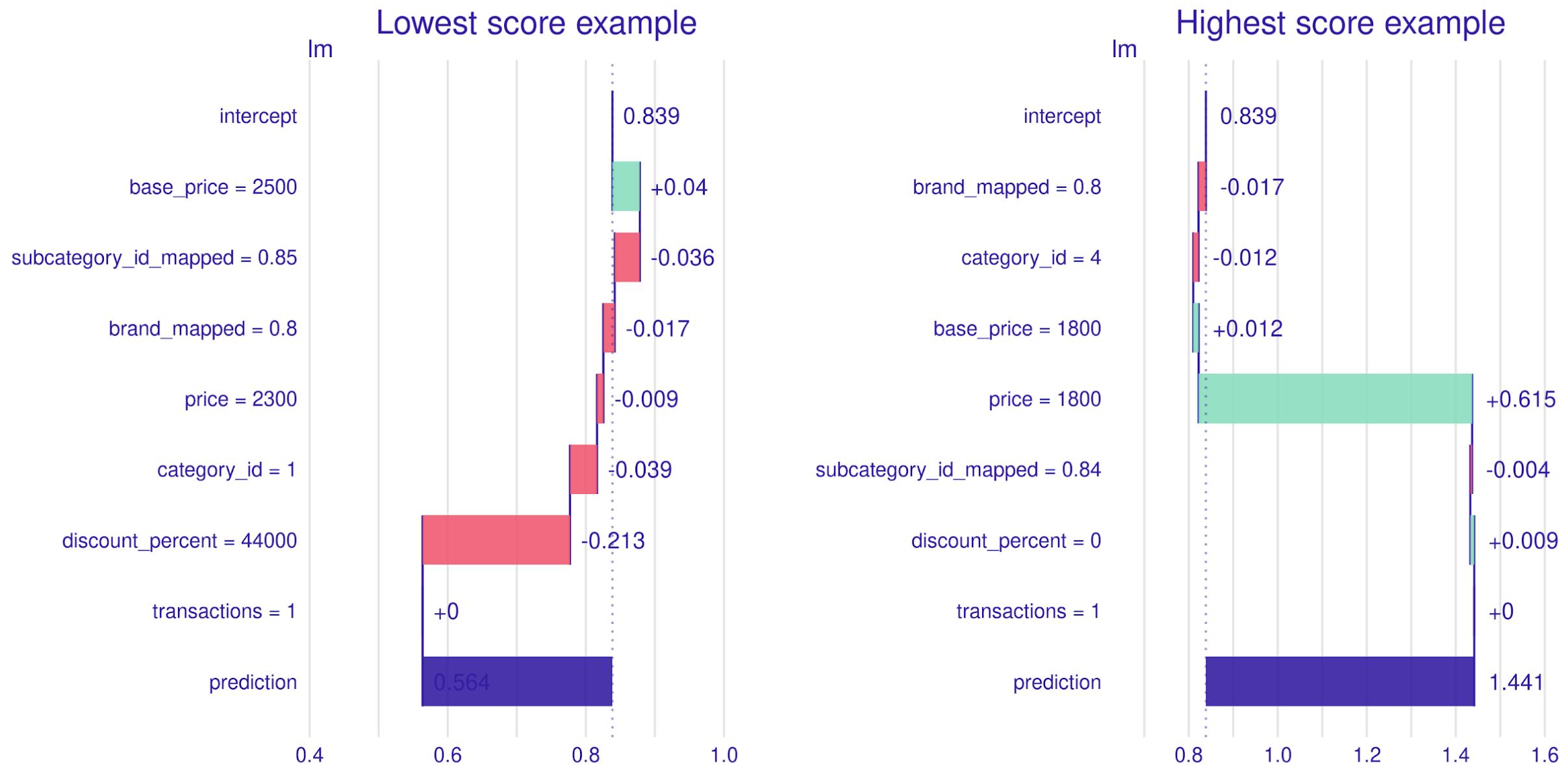


Ujemny loss-drop jest raczej wynikiem przypadkowym, zatem mamy tutaj jedną silną zmienną poza category tj. cenę

### Ceteris Paribus Profiles Highest score / random / lowest score in segment 3

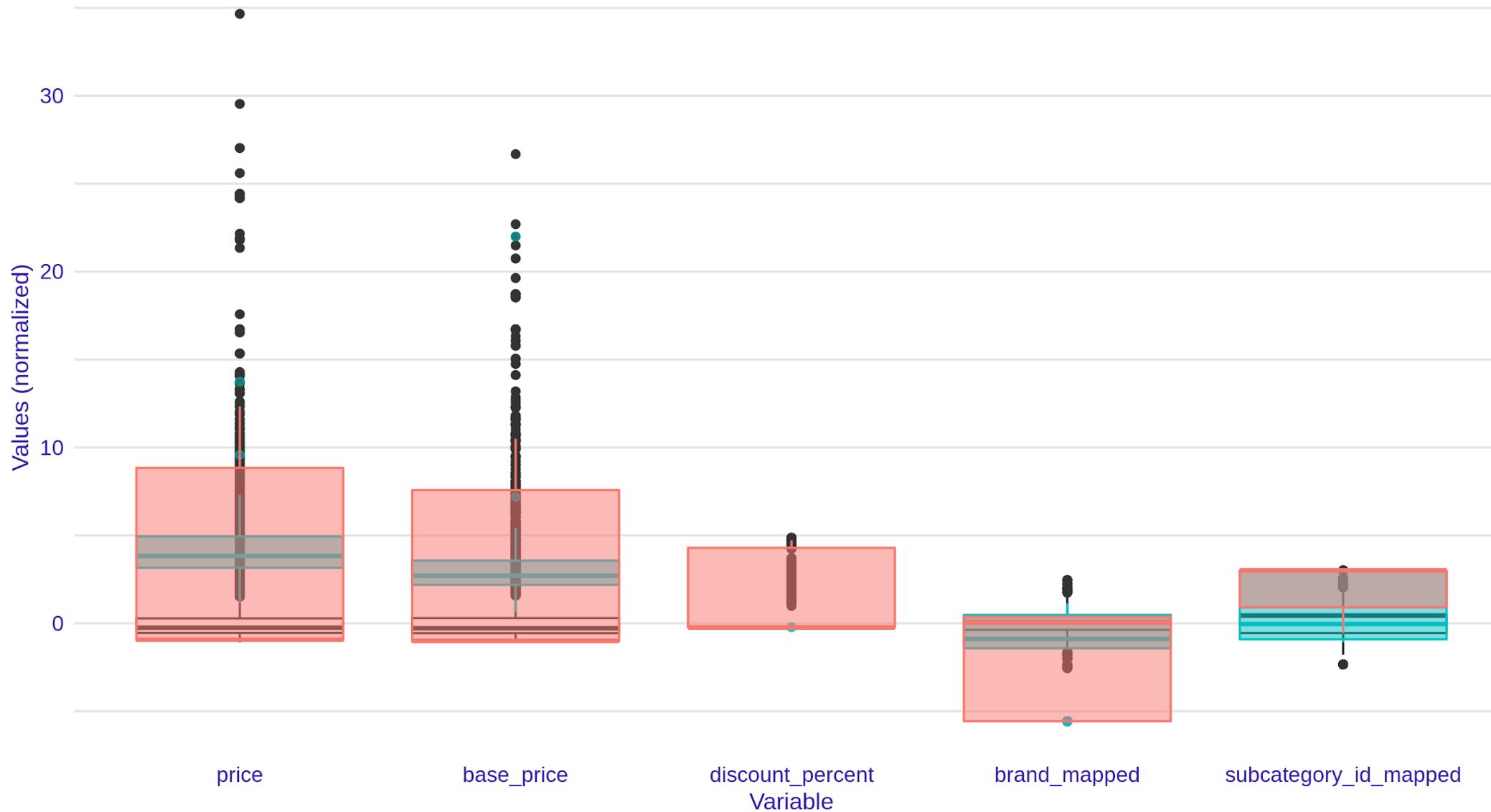


Widzimy jednak że ważność ceny jest silnie zależna od innych zmiennych.



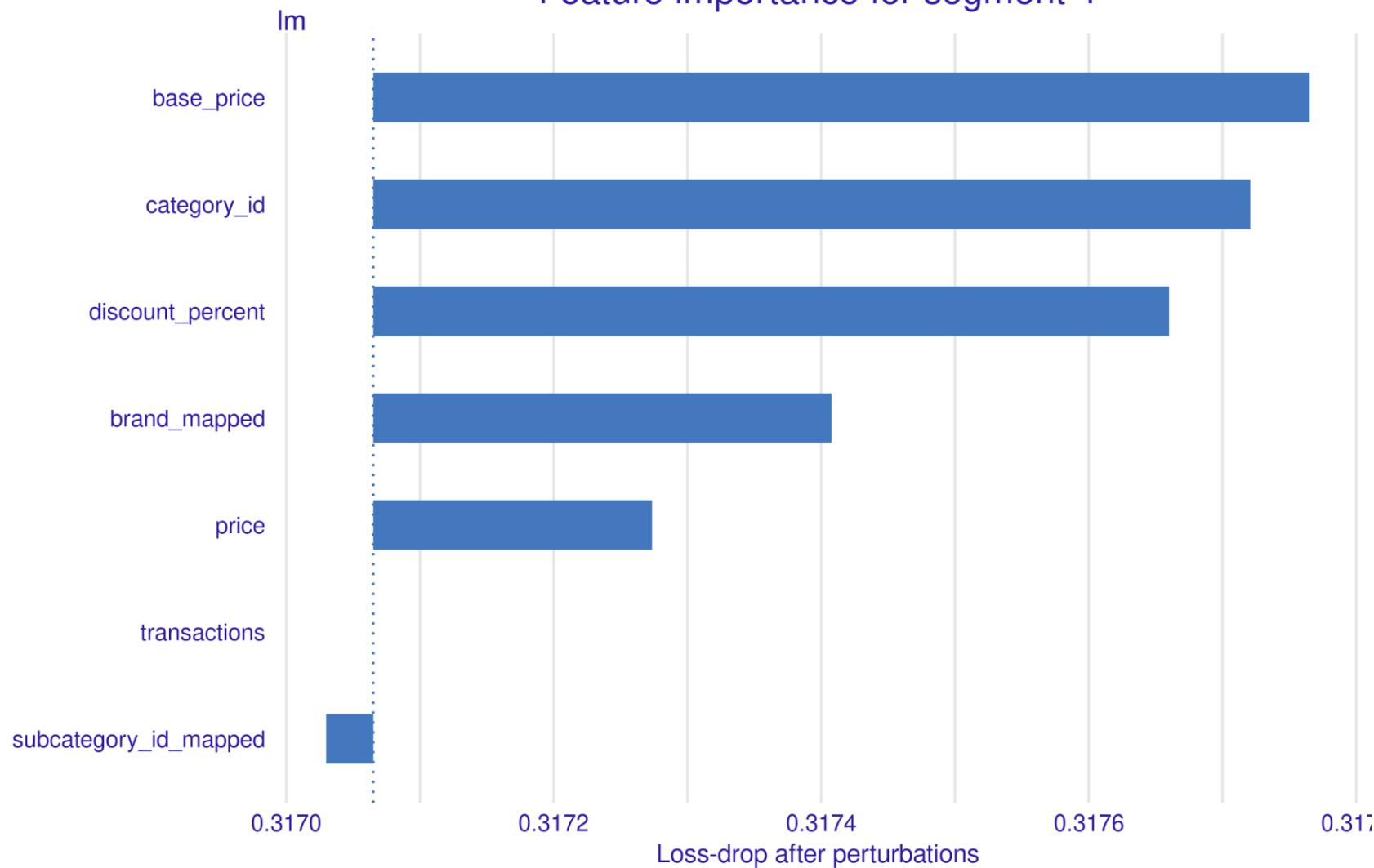
## Response box plot for segment 3

 Bottom 50  Top 50



Analiza per segment  
Segment 4

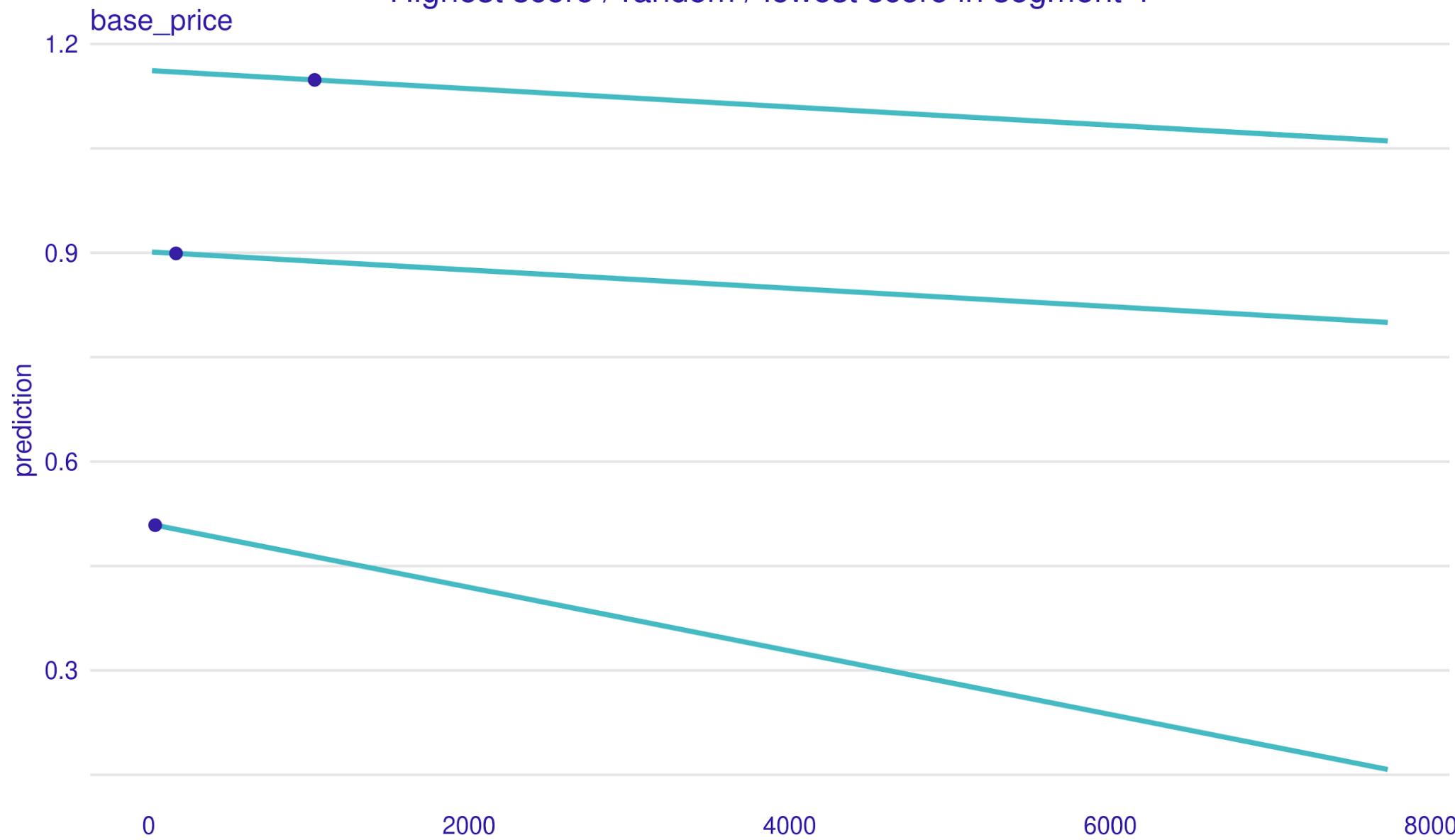
## Feature importance for segment 4

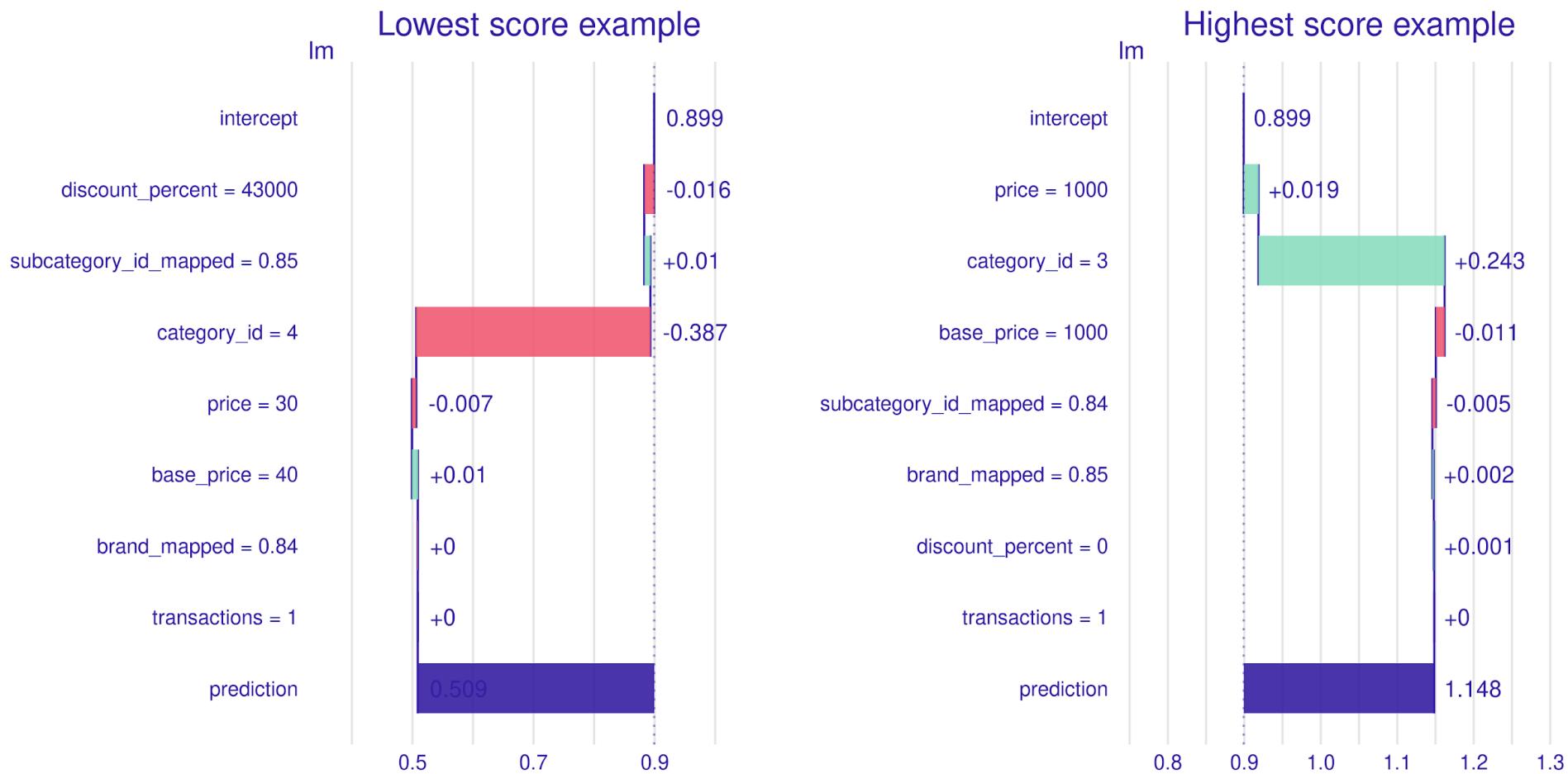


Przedostatni segment wykorzystuje wiedzę płynącą ze wszystkich zmiennych. Potwierdza to sens wydzielenie tej części mimo jej niewielkiego rozmiaru.

# Ceteris Paribus Profiles

## Highest score / random / lowest score in segment 4

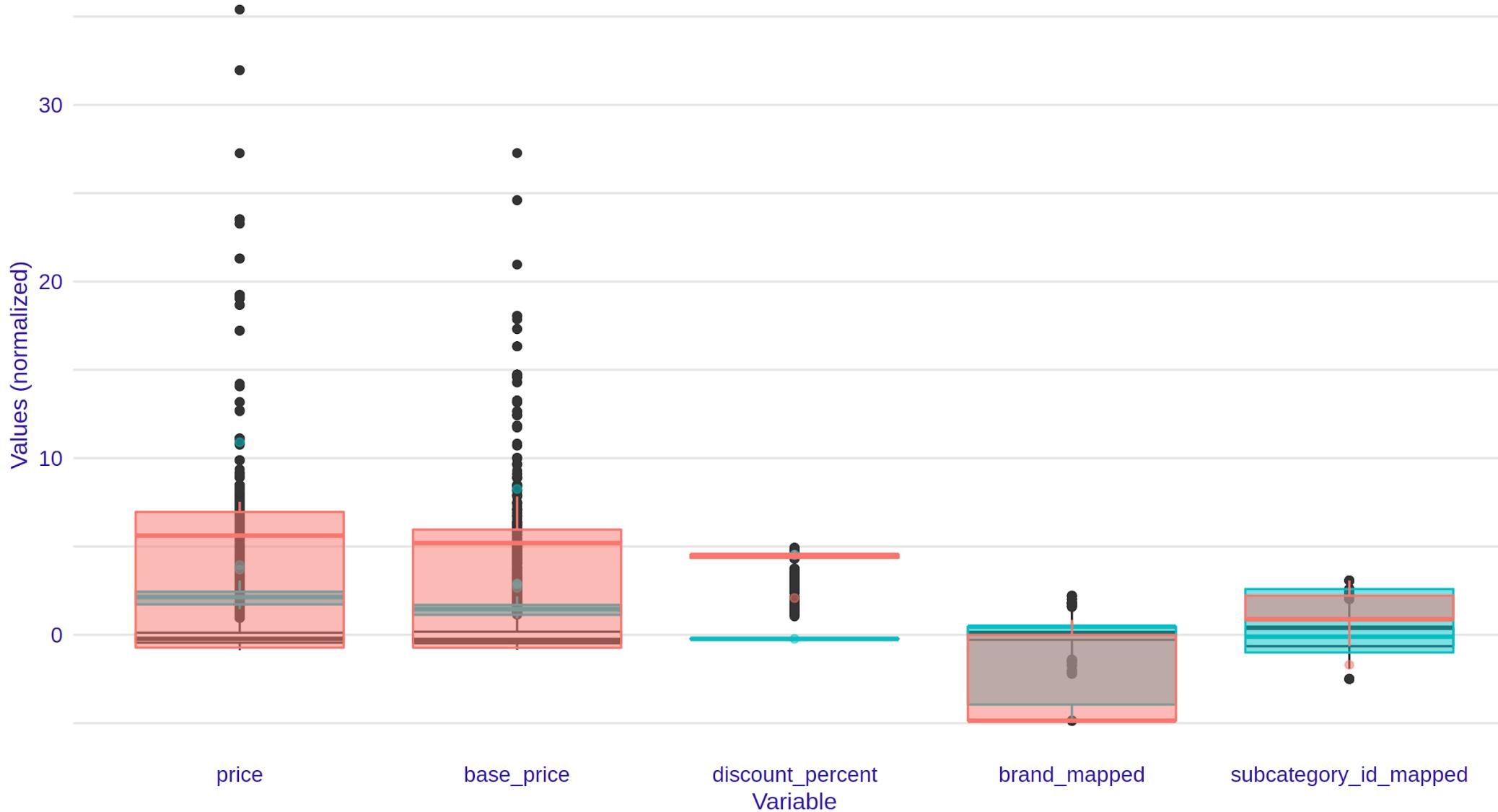




Pomimo że, prawie wszystkie zmienne miały wysoki feature importance to przy najlepiej i najgorzej ocenianym produkcie jedyną zmienną decydującą o tej pozycji była kategoria.

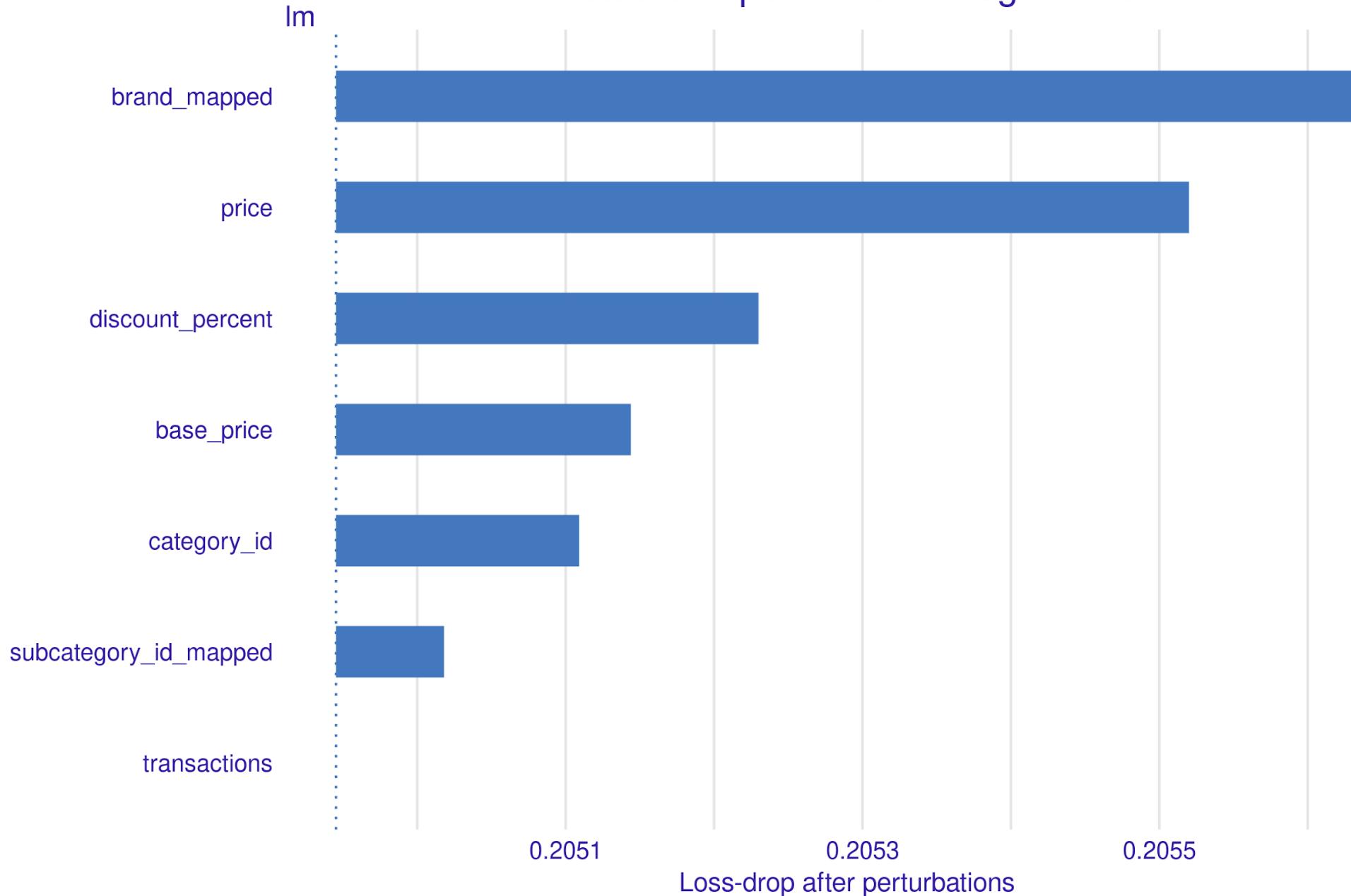
## Response box plot for segment 4

Bottom 50 Top 50



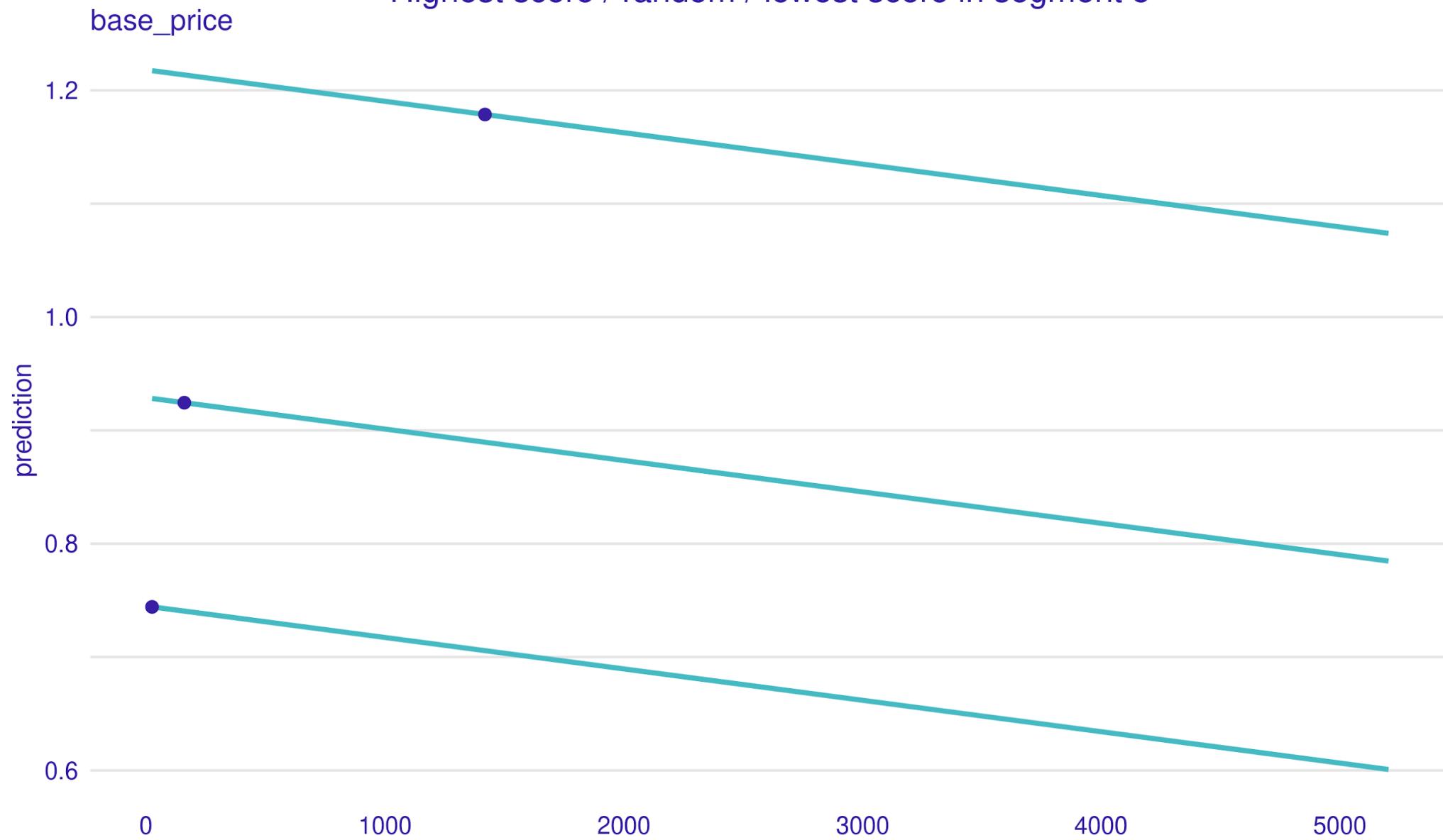
Analiza per segment  
Segment 5

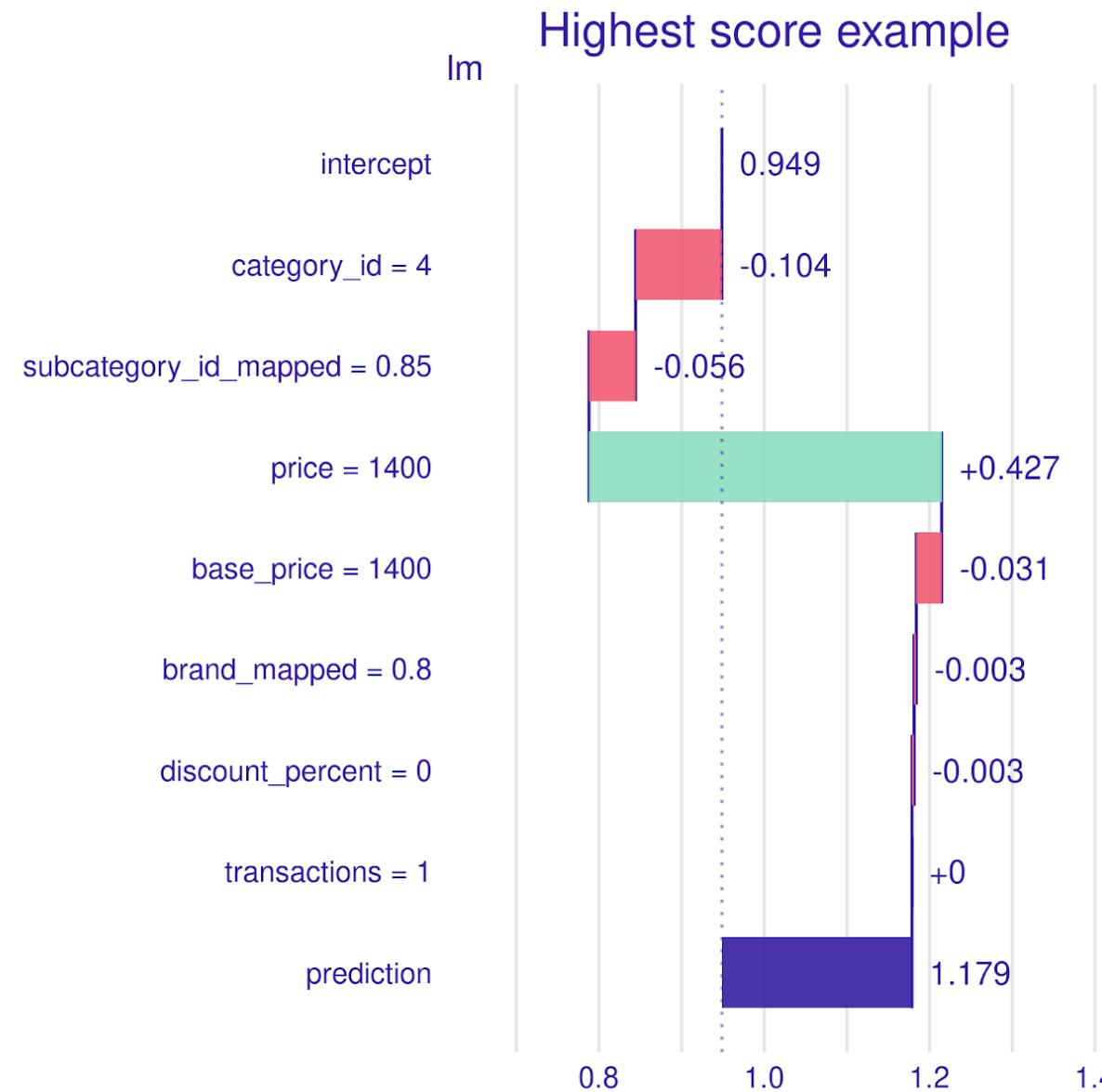
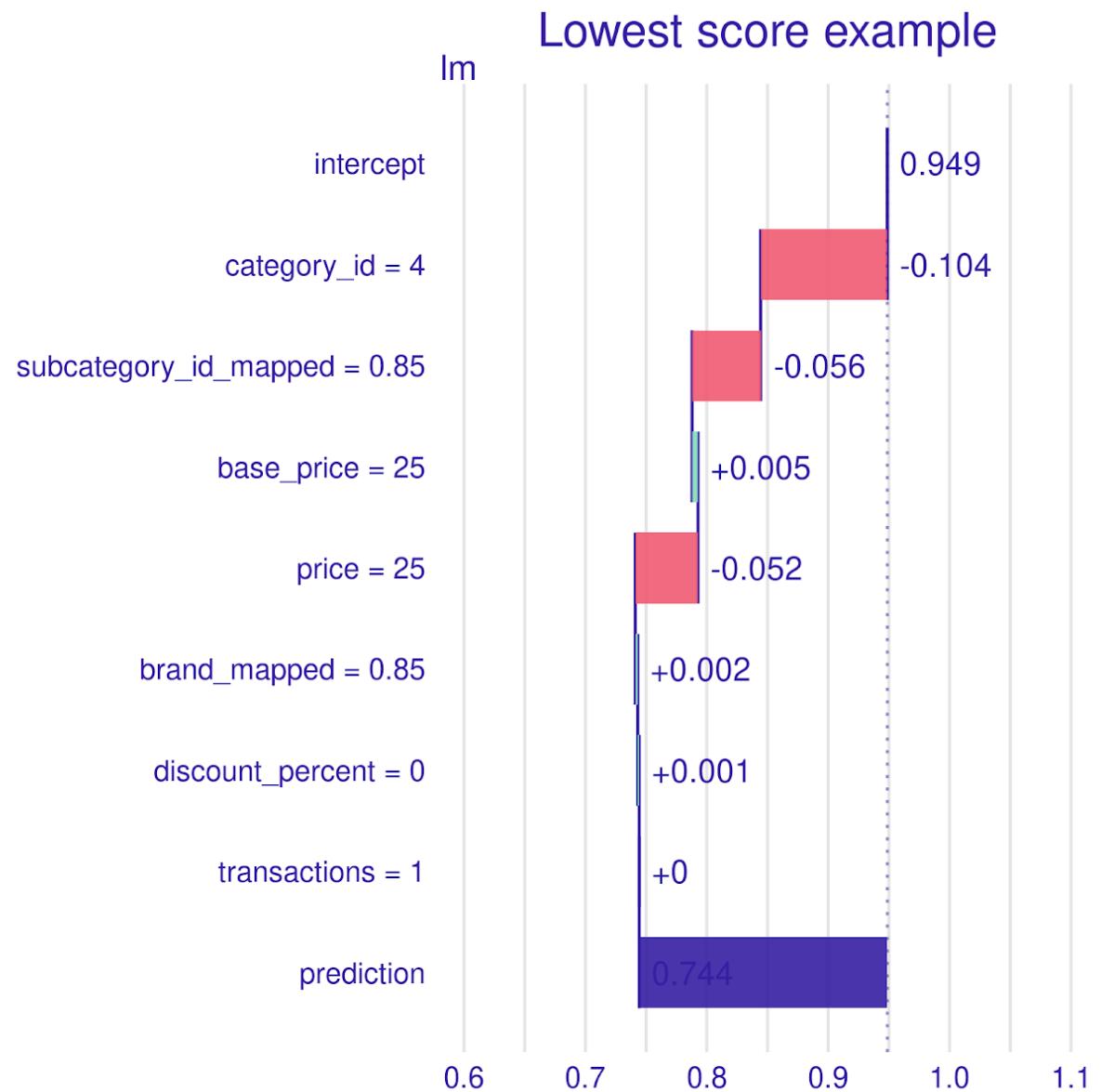
## Feature importance for segment 5



# Ceteris Paribus Profiles

## Highest score / random / lowest score in segment 5





## Response box plot for segment 5

Bottom 50 Top 50

