XGBoost feature importance analysis

В данном отчете риведены SHAP plots и краткий анализ для моделей градиентного бустинга на различных наборах данных. В последнем разделе приведены рекомендации по улучшению пайплайна и табличной модели.

Примечания

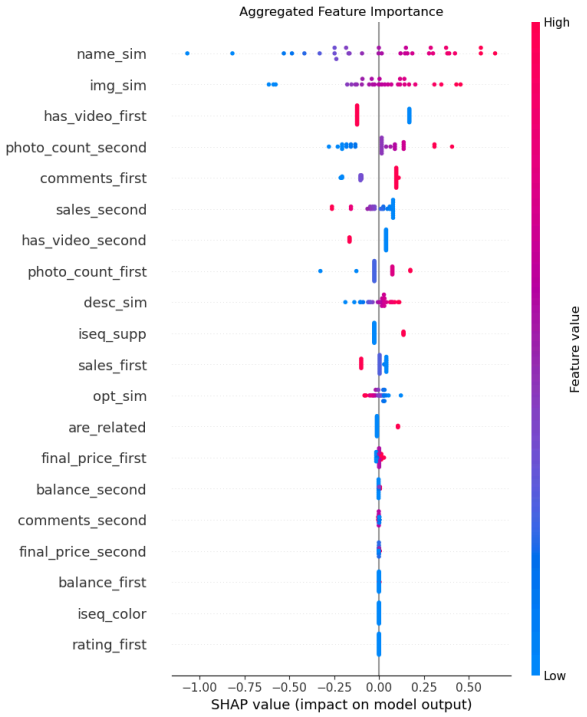
Набор данных `labeled.csv` (данные + файл разметки `sku_labeled_original_elena.csv)` - он же `WB_5k_paired` (5 тысяч пар товаров из разных категорий с разметкой конкурент/не конкурент)

1. Анализ признаков на WB-5k-paired

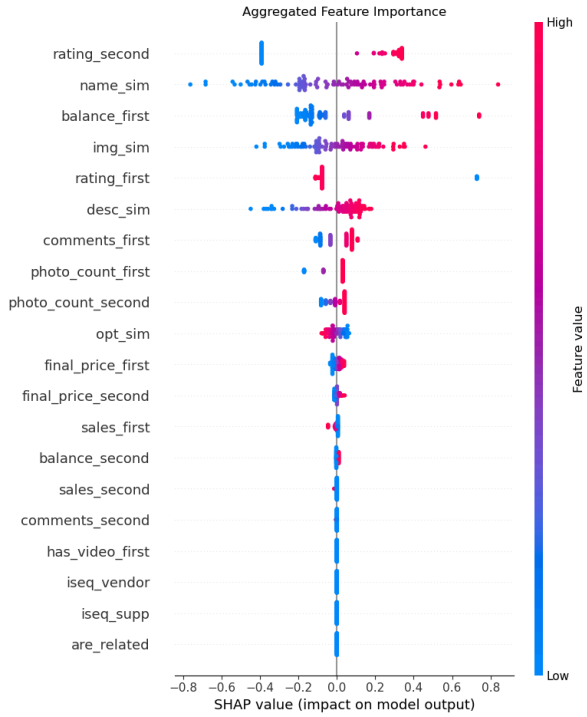
Используемые модели:

- model_params_big_test
- res_balanced_accuracy

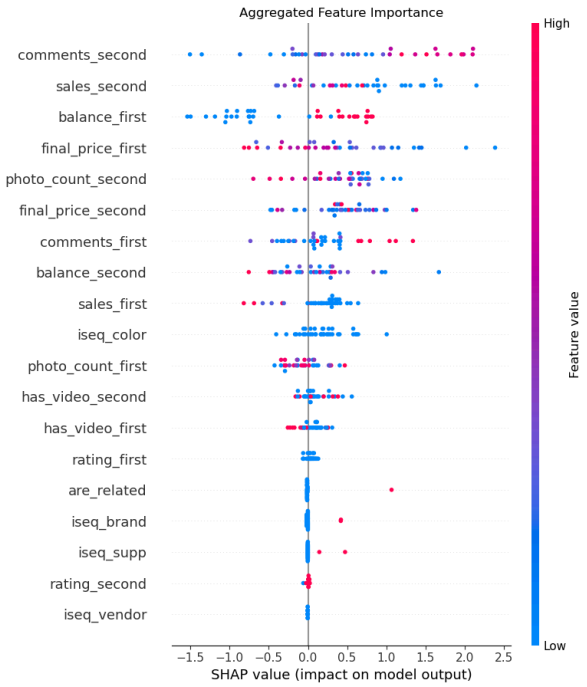
Используемый набор данных: `labeled.csv` (`data.csv` внутри чекпоинтов моделей). Модели были обучены на всех данных, без деления на train/val/test.



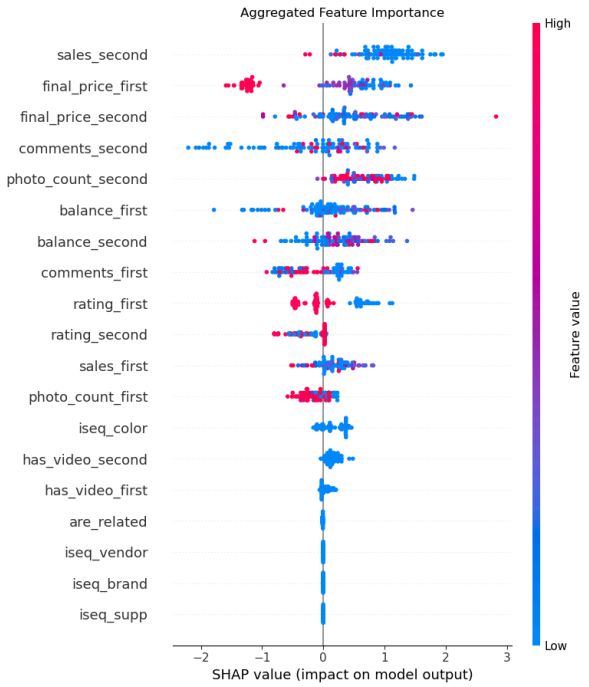
model_params_big_test - data.csv (regex 'карты')



model_params_big_test - tabular_OZ_geo_5500_top-50_query-23_nonquery-5539_embedded.csv



res_balanced_accuracy - data.csv (regex 'карты')



res_balanced_accuracy - tabular_OZ_geo_5500_top-50_query-23_nonquery-5539_embedded.csv

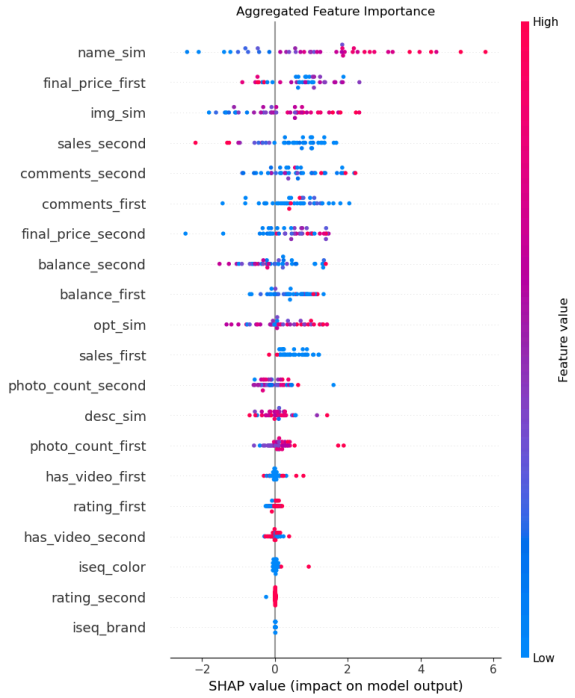
2. Анализ признаков на WB_5k_paired с выделением тестовой выборки

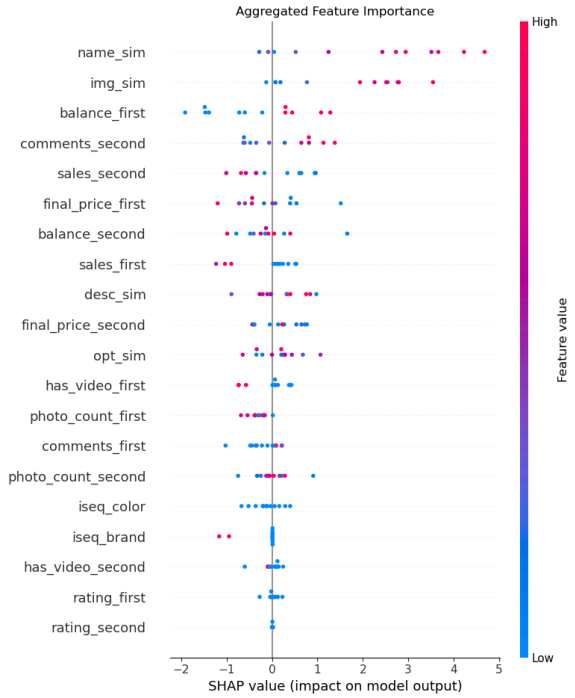
2.1 С скорами похожести товаров (sims=True)

Используемая модель: не сохранилась.

```
Accuracy:  0.8934
F1 Score:  0.8880
Precision: 0.9131
Recall:    0.8795

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.99      0.91        69
           1       0.98      0.77      0.86        53

    accuracy                           0.89       122
   macro avg       0.91      0.88      0.89       122
weighted avg       0.90      0.89      0.89       122
```

```
Accuracy:  0.8286
F1 Score:  0.6500
Precision: 0.6322
Recall:    0.6855

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.87      0.90        62
           1       0.33      0.50      0.40         8

    accuracy                           0.83        70
   macro avg       0.63      0.69      0.65        70
weighted avg       0.86      0.83      0.84        70
```





2.2 Со стратификацией по категориям

Используемая модель: `stratified_clusters=9`

category_id	category_name	category_size	accuracy	f1_score	precision	recall
0	Одежда женская (платья, юбки, блузки, кофты, б...	404	0.851485	0.838637	0.845745	0.833364
1	Одежда мужская (рубашки, футболки, брюки, шорт...	12	0.750000	0.733333	0.750000	0.728571
2	Одежда гимнастическая (гимнастическая форма, о...	56	0.839286	0.753786	0.784783	0.734347
5	Карта и путеводители (карты настенные, карты с...	65	0.876923	0.679803	0.750000	0.648810
6	Товары для уборки (перчатки резиновые, швабры,...	429	0.846154	0.825687	0.835188	0.818602
7	Товары для готовки (соевые соусы, ...)	3	1.000000	1.000000	1.000000	1.000000
8	Мебель (столы, стулья, диваны, кровати, ...)	11	1.000000	1.000000	1.000000	1.000000
9	Аксессуары для компьютеров (флешки с гравировк...	3	0.333333	0.250000	0.166667	0.500000
10	Всё остальное (игра настольная для детей, игру...	12	1.000000	1.000000	1.000000	1.000000

Метрики модели `stratified_clusters=9` на отложенном тесте (10%)

2.3 Со стратификацией по категориям, со скорами похожести обхектов

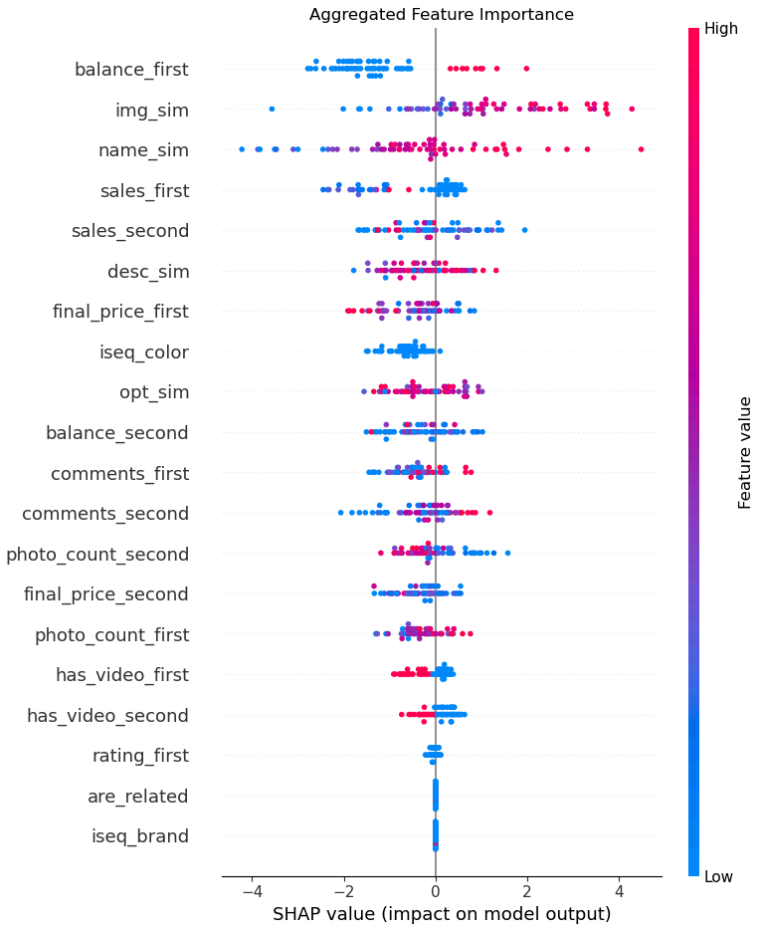
Используемая модель: `sims=True_stratified_clusters=9`

category_id	category_name	category_size	accuracy	f1_score	precision	recall
0	Одежда женская (платья, юбки, блузки, кофты, б...	404	0.861386	0.845919	0.866523	0.834594
1	Одежда мужская (рубашки, футболки, брюки, шорт...	12	0.583333	0.555556	0.562500	0.557143
2	Одежда гимнастическая (гимнастическая форма, о...	56	0.892857	0.849732	0.849732	0.849732
5	Карта и путеводители (карты настенные, карты с...	65	0.861538	0.695471	0.706140	0.686508
6	Товары для уборки (перчатки резиновые, швабры,...	429	0.834499	0.807758	0.829545	0.795314
7	Товары для готовки (соевые соусы, ...)	3	0.666667	0.666667	0.750000	0.750000
8	Мебель (столы, стулья, диваны, кровати, ...)	11	1.000000	1.000000	1.000000	1.000000
9	Аксессуары для компьютеров (флешки с гравировк...	3	0.666667	0.666667	0.750000	0.750000
10	Всё остальное (игра настольная для детей, игру...	12	0.916667	0.899160	0.944444	0.875000

Метрики модели `sims=True_stratified_clusters=9` на отложенном тесте (10%)

3. Анализ признаков на OZ_geo_5500 в разрезе FN, FP ошибок

Используемая модель: `model_params_big_test`



3.1 False Negatives (True=1, Predicted=0)

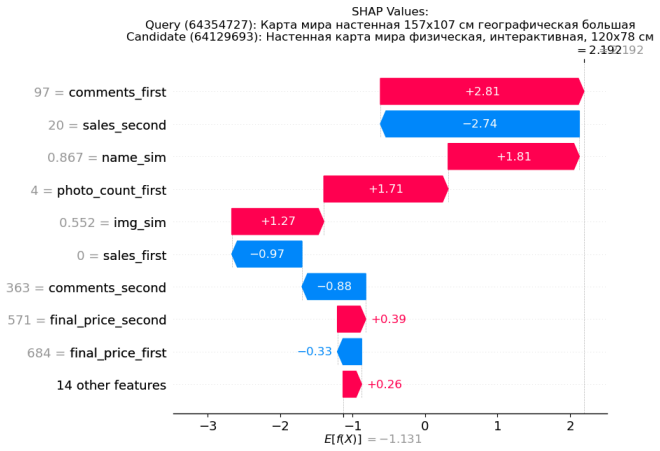
Очень похожи на матч по всем параметрам и имеют ненлувые продажи, но всё равно предсказаны как не-матч из-за:

- sales_second
- sales_first
- comments_second

При этом наибольшее положительное влияние оказали:

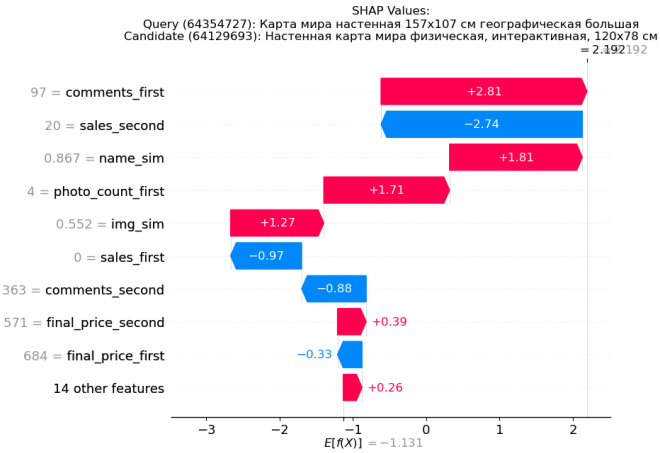
- comments_first (чем больше комментариев у query, тем выше шанс кандидатов)
- name_sim
- photo_count_first

```
Query and Candidate SKU fields:
------------------------------
               Query SKU  Candidate SKU
sku          178005436.00    54671906.00
final_price      1178.00         988.00
balance             0.00           0.00
sales               0.00           3.00
rating              5.00           5.00
comments           27.00        2075.00
name_sim            0.71           0.71
img_sim             0.68           0.68
desc_sim            0.88           0.88
opt_sim             0.93           0.93
```

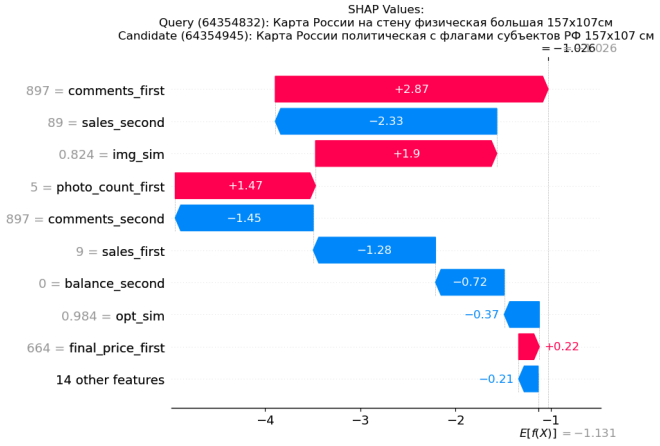


Разный размер, но обе физические

```
Query and Candidate SKU fields:
------------------------------
               Query SKU  Candidate SKU
sku          64354727.00    64129693.00
final_price      684.00         571.00
balance            0.00          62.00
sales              0.00          20.00
rating             5.00           5.00
comments          97.00         363.00
name_sim           0.87           0.87
img_sim            0.55           0.55
desc_sim           0.74           0.74
opt_sim            0.85           0.85
```




```
Query and Candidate SKU fields:
------------------------------
               Query SKU  Candidate SKU
sku          64354832.00    64354945.00
final_price      664.00         708.00
balance            0.00           0.00
sales              9.00          89.00
rating             5.00           5.00
comments         897.00         897.00
name_sim           0.73           0.73
img_sim            0.82           0.82
desc_sim           0.87           0.87
opt_sim            0.98           0.98
```



3.2 False Positives (True=0, Predicted=1)

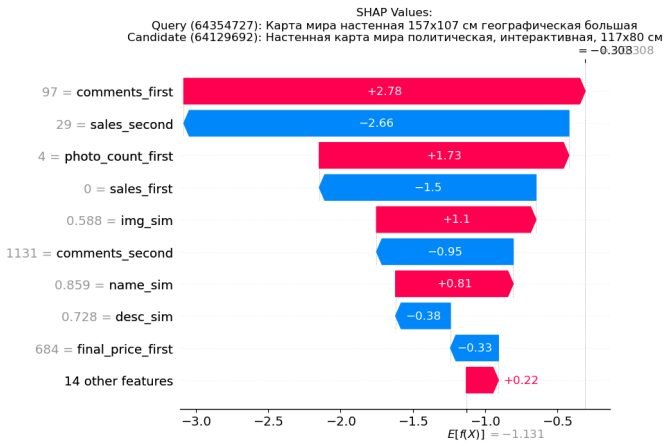
Очень похожи на матч по всем параметрам и имеют ненлувые продажи, но всё равно предсказаны как матч из-за:

- comments_first
- photo_count_first
- img_sim

Тем не менее, наибОльшее **негативное** влияние по-прежнему оказали:

- sales_second
- sales_first
- comments_second

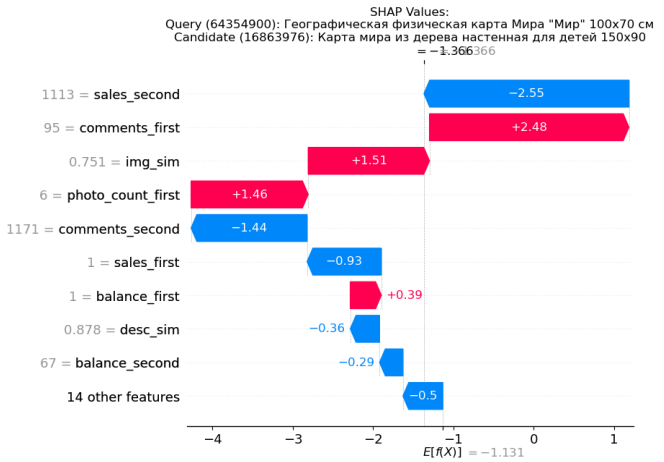
```
Query and Candidate SKU fields:
------------------------------
                Query SKU  Candidate SKU
sku          64354727.00    64129692.00
final_price      684.00         449.00
balance            0.00         178.00
sales              0.00          29.00
rating             5.00           5.00
comments          97.00        1131.00
name_sim           0.86           0.86
img_sim            0.59           0.59
desc_sim           0.73           0.73
opt_sim            0.82           0.82
```



Товары были бы (другой тип по атрибутам - одна физическая, другая физико-политическая)

Query and Candidate SKU fields:

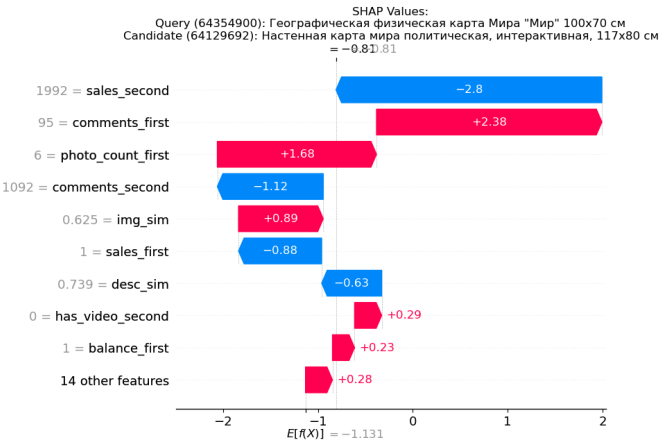
	Query SKU	Candidate SKU
sku	64354900.00	16863976.00
final_price	432.00	2525.00
balance	1.00	67.00
sales	1.00	1113.00
rating	5.00	5.00
comments	95.00	1171.00
name_sim	0.72	0.72
img_sim	0.75	0.75
desc_sim	0.88	0.88
opt_sim	0.80	0.80



Высокие похожести по эмбеддингам, хотя товары явно разные (другой тип по атрибутам - одна физическая, другая ДЕРЕВЯННАЯ)

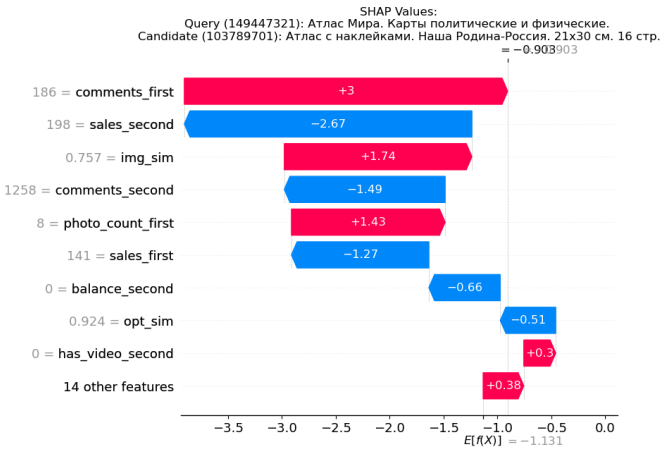
Query and Candidate SKU fields:

	Query SKU	Candidate SKU
sku	64354900.00	64129692.00
final_price	432.00	441.00
balance	1.00	309.00
sales	1.00	1992.00
rating	5.00	5.00
comments	95.00	1092.00
name_sim	0.78	0.78
img_sim	0.62	0.62
desc_sim	0.74	0.74
opt_sim	0.86	0.86



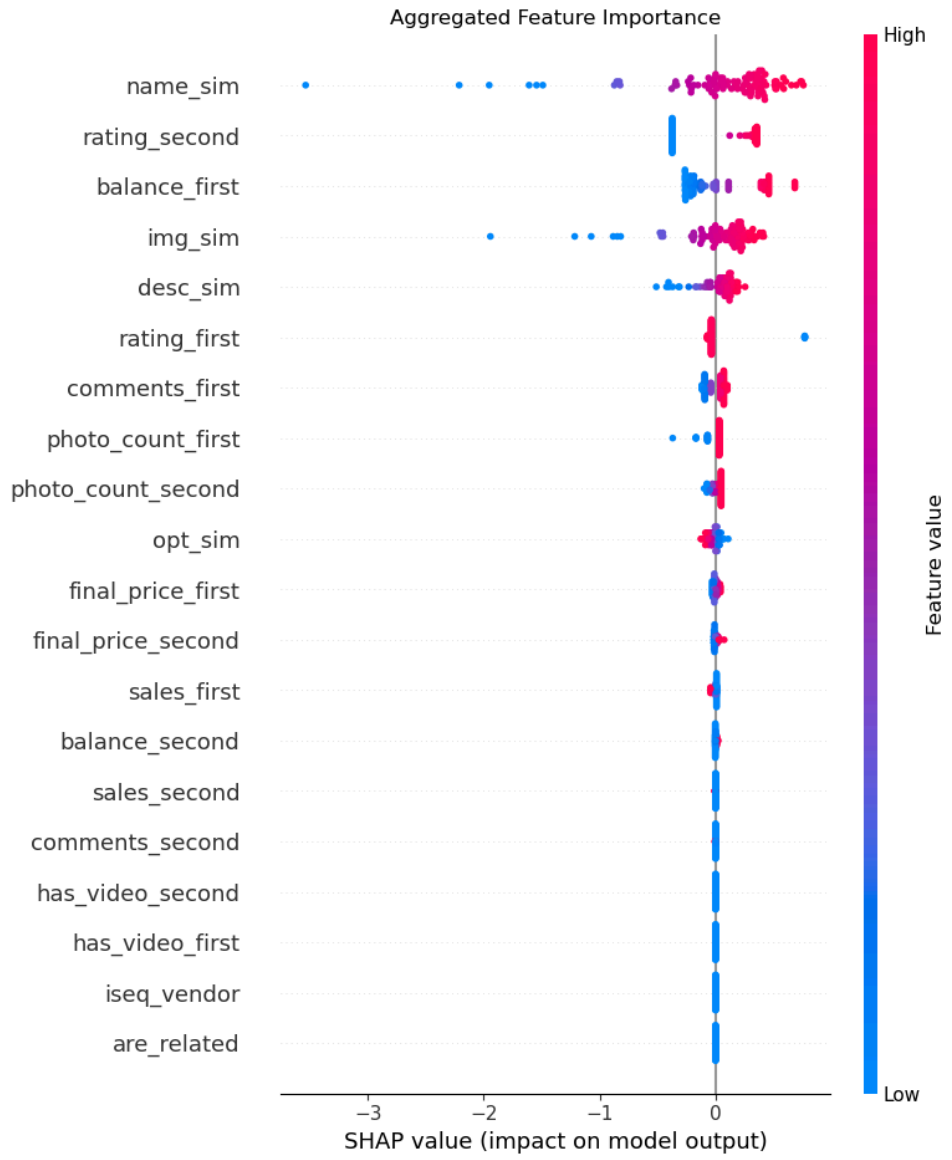
Товары разные, но сложно отличить по изображению или названию (другой тип по атрибутам - одна физическая, другая физико-политическая)

```
Query and Candidate SKU fields:
------------------------------
                Query SKU  Candidate SKU
sku         149447321.00   103789701.00
final_price       288.00         295.00
balance             0.00           0.00
sales             141.00         198.00
rating              5.00           5.00
comments          186.00        1258.00
name_sim            0.64           0.64
img_sim             0.76           0.76
desc_sim            0.88           0.88
opt_sim             0.92           0.92
```



Товары разные, но сложно отличить по изображению или названию (другой тип по атрибутам - одна карта мира, а другая карта России)

4. Анализ признаков на OZ_geo_5500 в разрезе FN, FP ошибок



4.1. Bad candidates

Кандидаты признаются матчами, несмотря на то что имеют совершенно различное содержание.

Предсказание смещается в положительную сторону из-за:

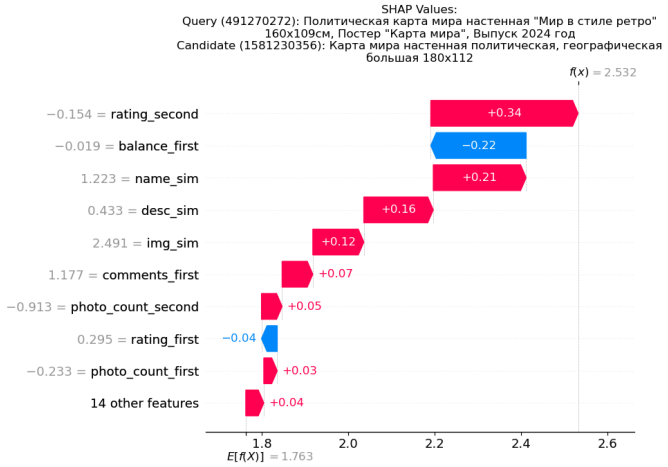
1. name_sim, img_sim (Высокие скоры схожести для различных товаров)
2. rating_second (кандидат имеет высокий рейтинг с **НИЗКИМ** кол-во отзывов)

3. balance_first (при высоком наличии целевого товара больше вероятность посчитать кого-либо как конкурента)
4. rating_first (при низком рейтинге целевого товара)

Query and Candidate SKU fields:

	Query SKU	Candidate SKU
sku	491270272.00	1581230356.00
final_price	1153.00	1196.00
balance	41.00	497.00
sales	9.00	0.00
rating	4.80	4.90
comments	1483.00	10.00
name_sim	0.82	0.82
img_sim	0.82	0.82
desc_sim	0.87	0.87
opt_sim	0.64	0.64

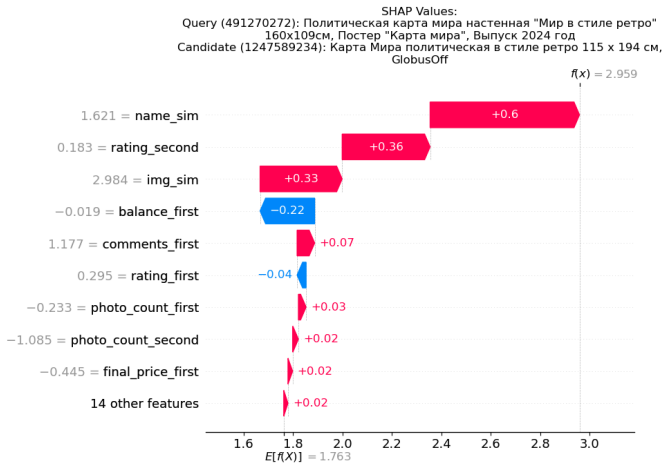
Высокие скоры схожести для различных товаров



Query and Candidate SKU fields:

	Query SKU	Candidate SKU
sku	491270272.00	1876314221.00
final_price	1153.00	668.00
balance	41.00	3.00
sales	9.00	0.00
rating	4.80	5.00
comments	1483.00	1.00
name_sim	0.83	0.83
img_sim	0.83	0.83
desc_sim	0.86	0.86
opt_sim	0.64	0.64

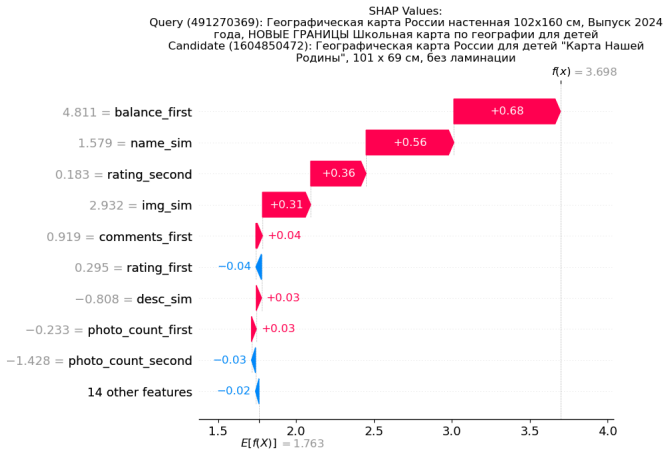
Высокие скоры схожести для различных товаров



Query and Candidate SKU fields:

	Query SKU	Candidate SKU
sku	491270369.00	1604850472.00
final_price	813.00	402.00
balance	899.00	50.00
sales	117.00	0.00
rating	4.80	5.00
comments	1257.00	4.00
name_sim	0.87	0.87
img_sim	0.87	0.87
desc_sim	0.78	0.78
opt_sim	0.68	0.68

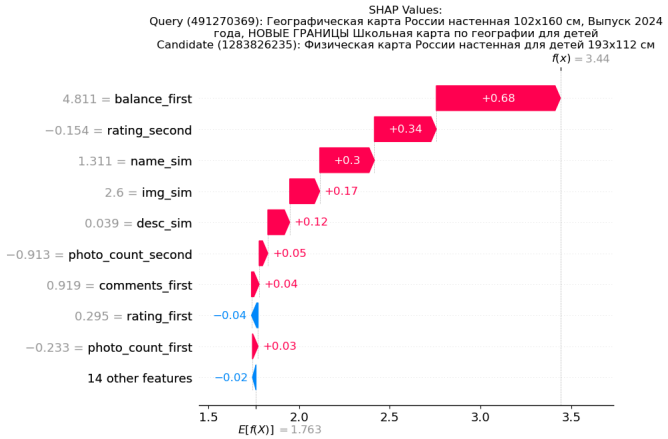
- Высокие скоры схожести для различных товаров;
- высокий баланс целевого товара



Query and Candidate SKU fields:

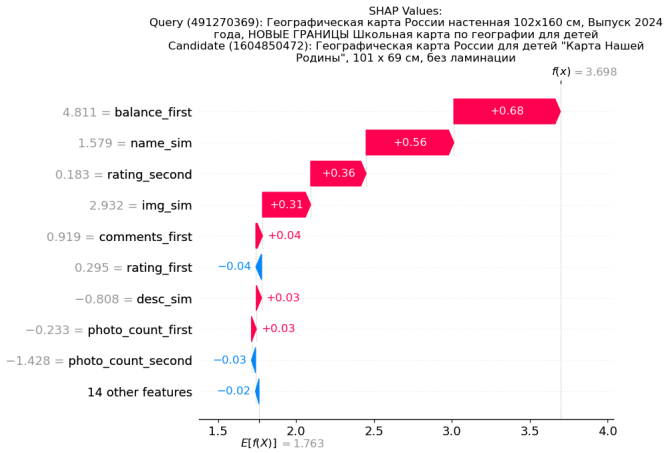
	Query SKU	Candidate SKU
sku	491270369.00	1283826235.00
final_price	813.00	1166.00
balance	899.00	691.00
sales	117.00	2.00
rating	4.80	4.90
comments	1257.00	90.00
name_sim	0.84	0.84
img_sim	0.84	0.84
desc_sim	0.84	0.84
opt_sim	0.70	0.70

- высокий баланс целевого товара
- различные товары по названию



Query and Candidate SKU fields:

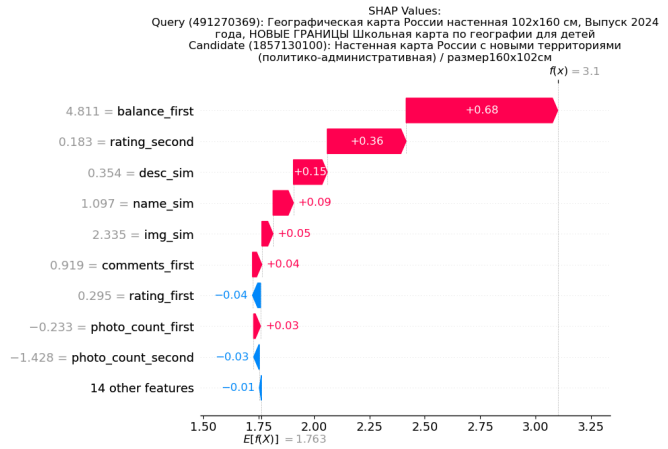
	Query SKU	Candidate SKU
sku	491270369.00	1604850472.00
final_price	813.00	402.00
balance	899.00	50.00
sales	117.00	0.00
rating	4.80	5.00
comments	1257.00	4.00
name_sim	0.87	0.87
img_sim	0.87	0.87
desc_sim	0.78	0.78
opt_sim	0.68	0.68



- высокий баланс целвого товара
- завышенный рейтинг
- различные товары по **КАТЕГОРИАЛЬНЫМ АТРИБУТАМ** с высокими скорами похожести

Query and Candidate SKU fields:

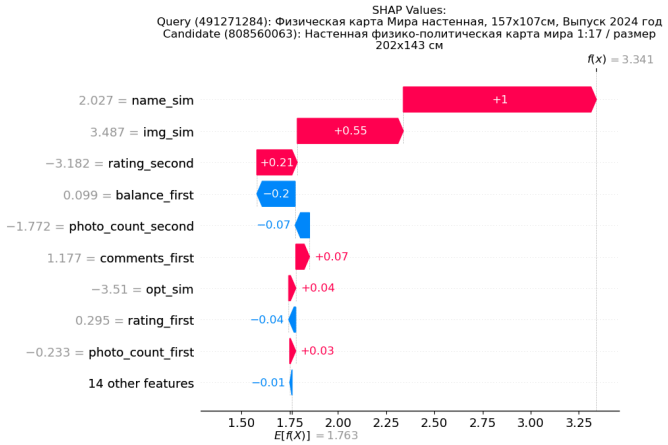
	Query SKU	Candidate SKU
sku	491270369.00	1857130100.00
final_price	813.00	1037.00
balance	899.00	14.00
sales	117.00	1.00
rating	4.80	5.00
comments	1257.00	62.00
name_sim	0.81	0.81
img_sim	0.81	0.81
desc_sim	0.86	0.86
opt_sim	0.66	0.66



- различные товары по **ЧИСЛОВЫМ АТРИБУТАМ** с высокими скорами похожести

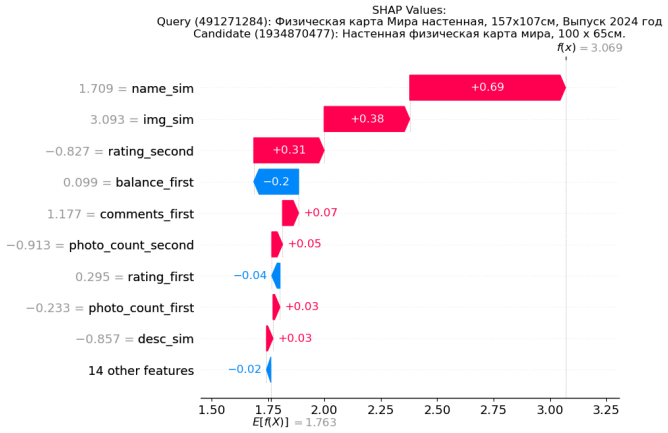
```
Query and Candidate SKU fields:
----------------------------
                Query SKU  Candidate SKU
sku          491271284.00   808560063.00
final_price       821.00        2411.00
balance            62.00           5.00
sales              24.00           0.00
rating              4.80           4.00
comments         1483.00          22.00
name_sim            0.94           0.94
img_sim             0.94           0.94
desc_sim            0.74           0.74
opt_sim             0.61           0.61
```

- различные товары по **ЧИСЛОВЫМ АТРИБУТАМ** с высокими скорами похожести

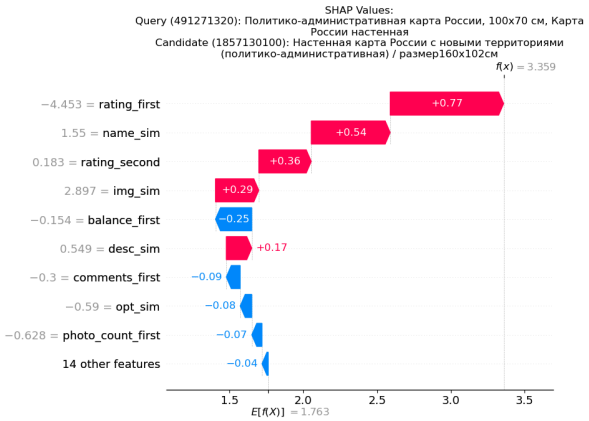



```
Query and Candidate SKU fields:
----------------------------
                Query SKU  Candidate SKU
sku          491271284.00  1934870477.00
final_price       821.00         383.00
balance            62.00          76.00
sales              24.00           2.00
rating              4.80           4.70
comments         1483.00         198.00
name_sim            0.89           0.89
img_sim             0.89           0.89
desc_sim            0.78           0.78
opt_sim             0.69           0.69
```

- различные товары по **КАТЕГОРИАЛЬНЫМ АТРИБУТАМ** с высокими скорами похожести



```
Query and Candidate SKU fields:
------------------------------
                Query SKU  Candidate SKU
sku          491271320.00  1857130100.00
final_price       522.00        1037.00
balance            17.00          14.00
sales               0.00           1.00
rating              0.00           5.00
comments          185.00          62.00
name_sim            0.87           0.87
img_sim             0.87           0.87
desc_sim            0.88           0.88
opt_sim             0.82           0.82
```



- различные товары по **ЧИСЛОВЫМ АТРИБУТАМ** с высокими скорами похожести

4.2. Bad predictions over bad candidates

Кандидаты признаются матчами, несмотря на низкие продажи, баланс.

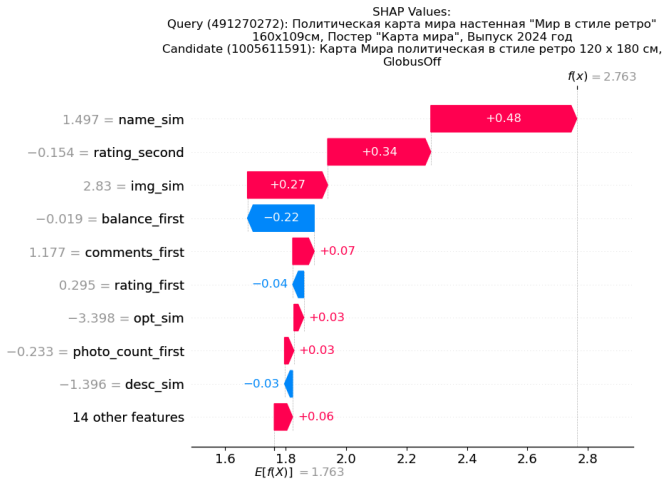
Предсказание смещается в положительную сторону из-за:

- name_sim (высокие скоры для непохожих товаров)
- rating_second
  - (кандидат имеет высокий рейтинг с **НИЗКИМ** кол-во отзывов)
  - (кандидат имеет высокий рейтинг со **СРЕДНИМ** кол-во отзывов)
- img_sim (высокие скоры для непохожих товаров)
- rating_first (целевой товар имеет низкий рейтинг)

Query and Candidate SKU fields:

	Query SKU	Candidate SKU
sku	491270272.00	1005611591.00
final_price	1153.00	4049.00
balance	41.00	3.00
sales	9.00	0.00
rating	4.80	4.90
comments	1483.00	67.00
name_sim	0.86	0.86
img_sim	0.86	0.86
desc_sim	0.74	0.74
opt_sim	0.62	0.62

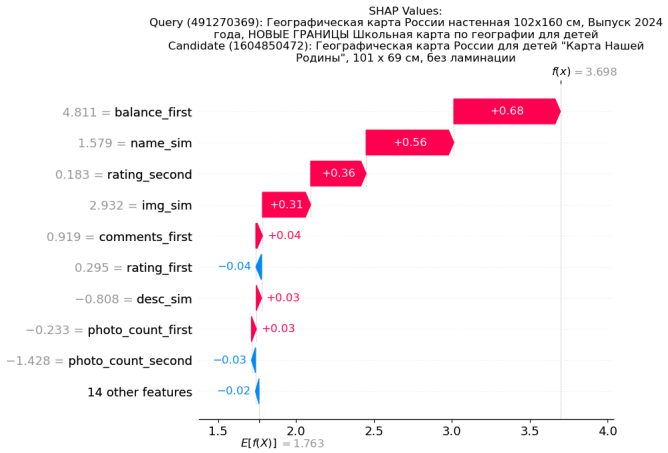
- высокие скоры похожести для совершенно различных товаров
- высокий рейтинг со средним кол-вом комментариев



Query and Candidate SKU fields:

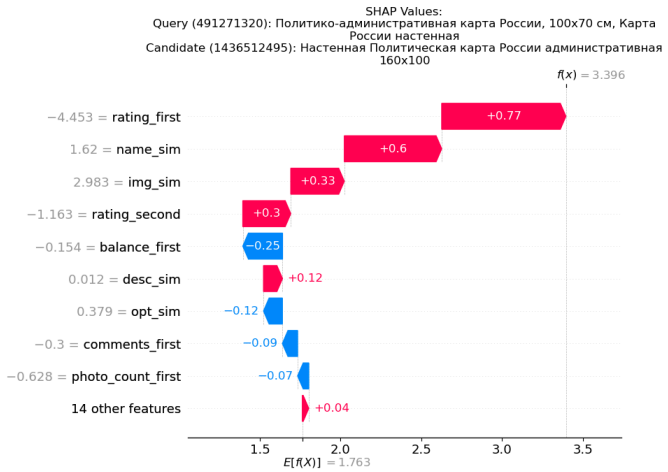
	Query SKU	Candidate SKU
sku	491270369.00	1604850472.00
final_price	813.00	402.00
balance	899.00	50.00
sales	117.00	0.00
rating	4.80	5.00
comments	1257.00	4.00
name_sim	0.87	0.87
img_sim	0.87	0.87
desc_sim	0.78	0.78
opt_sim	0.68	0.68

- баланс целевого
- высокий рейтинг с низким кол-вом комментариев



Query and Candidate SKU fields:

	Query SKU	Candidate SKU
sku	491271320.00	1436512495.00
final_price	522.00	555.00
balance	17.00	1000.00
sales	0.00	0.00
rating	0.00	4.60
comments	185.00	165.00
name_sim	0.88	0.88
img_sim	0.88	0.88
desc_sim	0.84	0.84
opt_sim	0.89	0.89



5. Рекомендации по улучшению модели

Таким образом, для улучшения определения кандидатов, необходимо:

1. На вход табличной модели подавать скор похожести, адекватно отражающий различие товаров внутри категории. Для этого нужно файн-тюнить модель различения товаров по содержанию для каждой категории (необходимо больше примеров на категорию - в WB_5k_paired максимум 300 пар на категорию).
2. Убрать неэффективные признаки:
    a. Не подавать *rating_second* в табличную модель, либо нормализовать рейтинг в соответствии с кол-во оценок
    b. Не подавать *balance_first* в табличную модель, либо составить достаточно примеров, где высокий баланс у целевого товара, но кандидаты не являются конкурентами (*label=0)*
3. Добавить относительные признаки типа:
    a. Отношение цен товаров (price_first / price_second)
    b. Отношение наличия товаров товаров (balance_first / balance_second)
    c. Отношение продаж товаров
    d. И т.д.