ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

OZON

OZONtech

КОМАНДА «ВОСТОК-1»

ЗАДАЧА 5

ПОИСК ОДИНАКОВЫХ ТОВАРОВ НА МАРКЕТПЛЕЙСЕ

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Состав команды

Владимир Матросов

- Data Scientist
- @morris_day73

Екатерина Куликова

- Data Scientist
- @EkaterinaTretia

Игорь Дерябин

- Data Scientist
- @Deryabin_Igor

Илья Соловьев

- Data Scientist
- @iLya_s_ds

Роман Глазов

- Data Scientist
- @happosaj

Мы – команда выпускников Яндекс.Практикума по профессии «специалист по Data Science». Нам нравится компания OZON и продукт, который она создает. Мы хотели прикоснуться к реальной рабочей задаче сразу после обучения и целенаправленно выбрали задачу этой компании.

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Постановка задачи

Разработать ML-модель, способную определить идентичность товаров по названиям, атрибутам и изображениям.

Модель должна находить среди пар-кандидатов как можно больше одинаковых товаров с высокой точностью.

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

План работы

1. • Составить гипотезы
2. • Подготовить данные, добавить дополнительные признаки, преобразовать имеющиеся
3. • Обучить модели, сравнить значения полученных метрик
4. • Объединить удачные идеи в финальное решение
5. • Получить финальную метрику, сделать выводы

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Гипотезы

1

Векторизация текстовых признаков

- Перевод текстовых признаков в векторный формат и вычисление tf-idf, позволит модели более точно находить матчи среди пар товаров.

2

Новые эмбеддинги текстовых признаков BERT

- Добавление новых, более полных эмбеддингов с помощью BERT, позволит модели более точно находить матчи среди пар товаров.

3

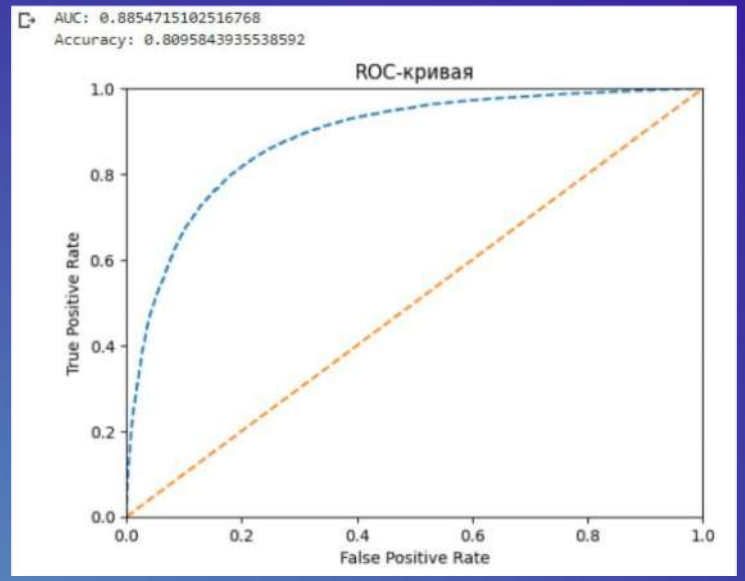
Дополнительные признаки, косинусное и Евклидово расстояние

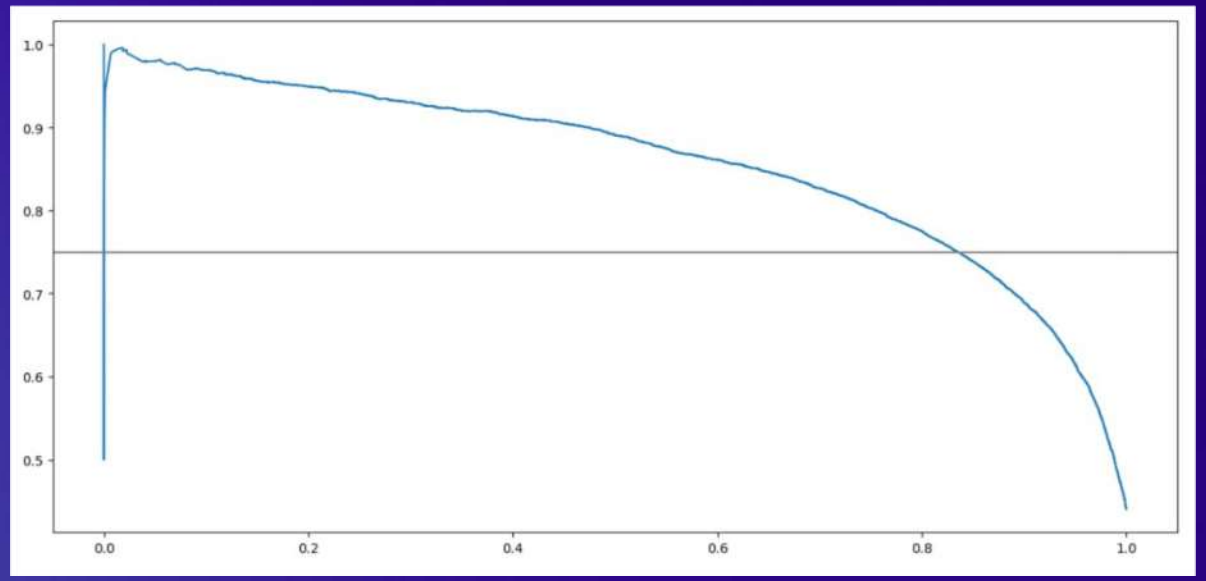
- Создание признаков, косинусное и евклидово расстояние между текстовыми признаками, позволит модели более точно находить матчи среди пар товаров.

Проверка гипотез, векторизация

Векторизация текстовых признаков

В ходе работы, выяснили, что на процесс векторизации признаков уходит много времени и также значение метрики не самое высокое. От этой идеи решено отказаться.





Проверка гипотез, дополнительные эмбеддинги

Дополнительные эмбеддинги текстовых признаков

	variantid	name_embedding
0	51195767	[-0.52616537, 0.5570387, 0.19487007, -0.363480...
1	53565809	[-0.64783597, -0.1698648, 0.4192899, -0.225463...
2	56763357	[-0.45637688, -0.0019165785, -0.2057866, -0.43...
3	56961772	[-0.61946654, 0.18796882, 0.07321368, -0.21601...
4	61054740	[-0.43500108, 0.26390466, -0.14602755, -0.2979...
5	65143063	[-0.44750938, 0.09084959, -0.011524273, -0.308...
6	66777498	[-0.6774125, 0.27191123, 0.13661727, -0.341242...
7	68392618	[-0.4427373, 0.057696387, 0.20258707, -0.62042...
8	73780268	[-0.5938448, 0.22875759, -0.073359095, -0.3721...
9	77646583	[-0.75335115, 0.18644702, 0.03092286, -0.33731...

В ходе работы получили дополнительные эмбеддинги с помощью модели BERT, где каждая строка длинной в 768 объектов, на такой обучающей выборки

Это было удачное решение и мы применили его в дальнейшем, в создании финальной модели.

Проверка гипотез, косинусное и Евклидово расстояние

n_pic_dist	euclidean_color_dist	cosine_color_dist	euclidean_name_bert_dist	cosine_name_bert_dist	euclidean_name_embedding_dist	cosine_name_embedding_dist
0.000387	0.000000	0.0	0.338458	0.003606	1.816846	0.007682
0.000456	0.000000	0.0	0.475419	0.007311	0.527725	0.000675

main_pic_dist_0_perc	main_pic_dist_25_perc	main_pic_dist_50_perc	euclidean_main_pic_dist	cosine_main_pic_dist	euclidean_color_dist	cosine_color_dist	euclide
0.259265	0.259265	0.259265	0.259265	0.000387	0.000000	0.0	
0.282023	0.282023	0.282023	0.282023	0.000456	0.000000	0.0	

В ходе исследования, вычислили различные виды расстояний, такие как косинусное и синусное, а также коэффициент Жаккара между различными парами признаков. Такие дополнительные признаки позволили модели более точно находить матчи среди пар товаров. Это было удачное решение и мы применили его в дальнейшем, в создании финальной модели.

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

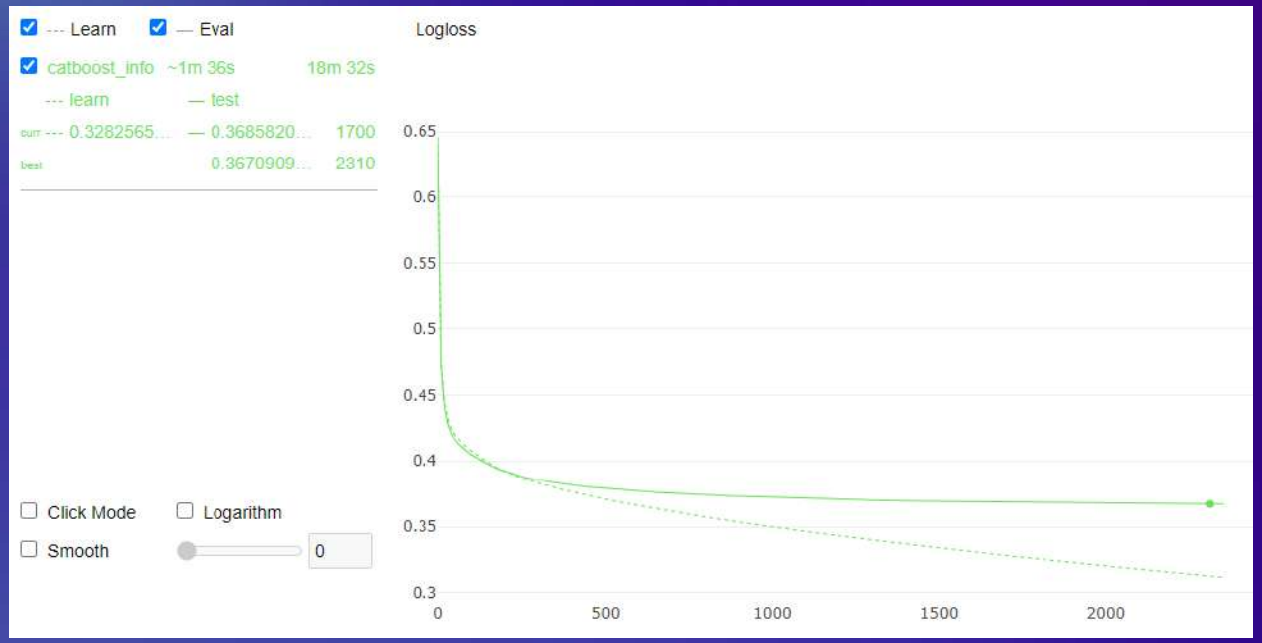
Особенности

	Feature Id	Importances
0	name2	16.578146
1	name1	15.119951
2	characteristic_attributes_mapping2	8.885653
3	name_jaccard_similarity	8.869428
4	characteristic_attributes_mapping1	8.533828
5	attr_dist	5.897497
6	cosine_name_embedding_dist	4.272765
7	euclidean_name_bert_dist	3.353865
8	euclidean_name_embedding_dist	2.877716
9	name_std_diff	2.257062
10	main_pic_std_diff	2.033271

После обучения моделей, мы получили рейтинг важности признаков. Помимо имен, верх списка попали такие признаки как коэффициент Жаккара между названиями, косинусное и евклидово расстояние между признаками.

Создание этих признаков ранее позволило модели более точно находить матчи среди пар товаров.

Финальная модель



Для обучения финальной модели мы применили удачные идеи и наработки с предыдущих этапов, такие как:

- Дополнительные эмбеддинги текстовых признаков
- Дополнительные признаки с разными типами расстояний
- Коэффициент Жаккара
- Преобразование в цифровой вид колонок с цветом
- манхэттенское расстояние между эмбеддингами

- расстояние Левенштейна для колонок с названиями и статистические характеристиками разности товаров, такие как разность между эмбеддингами, среднее значение разности, медиану разности, стандартное отклонение этой разности.

В итоге мы получили промежуточные метрики:

- PROC = 0.6573
- Accuracy = 0.8398
- Precision = 0.8116
- AUC = 0.9120
- F1 = 0.8201

А также метрику PR-AUC на публичном лидерборде:

Дата	Pr_Auc_Macro
29.05.2023, 2:47	0.29038