E-COMMERCE RECSYS
HSE CHECKPOINT № 5

СОДЕРЖАНИЕ

ПОСТАНОВКА ЗАДАЧИ

Требуется разработать рекомендательную систему на базе классического ML, которая будет способна предлагать релевантные товары клиентам магазина.

Состав команды

Роль	ФИО
Исполнитель	Поскребышев Сергей
Куратор	Ижеев Сергей

ОПИСАНИЕ ДАННЫХ

С платформы kaggle получили 2 датафрейма:
• products.csv - товары с их характеристиками
• transactions.csv - транзакции покупателей

products.csv (49 688 строк):

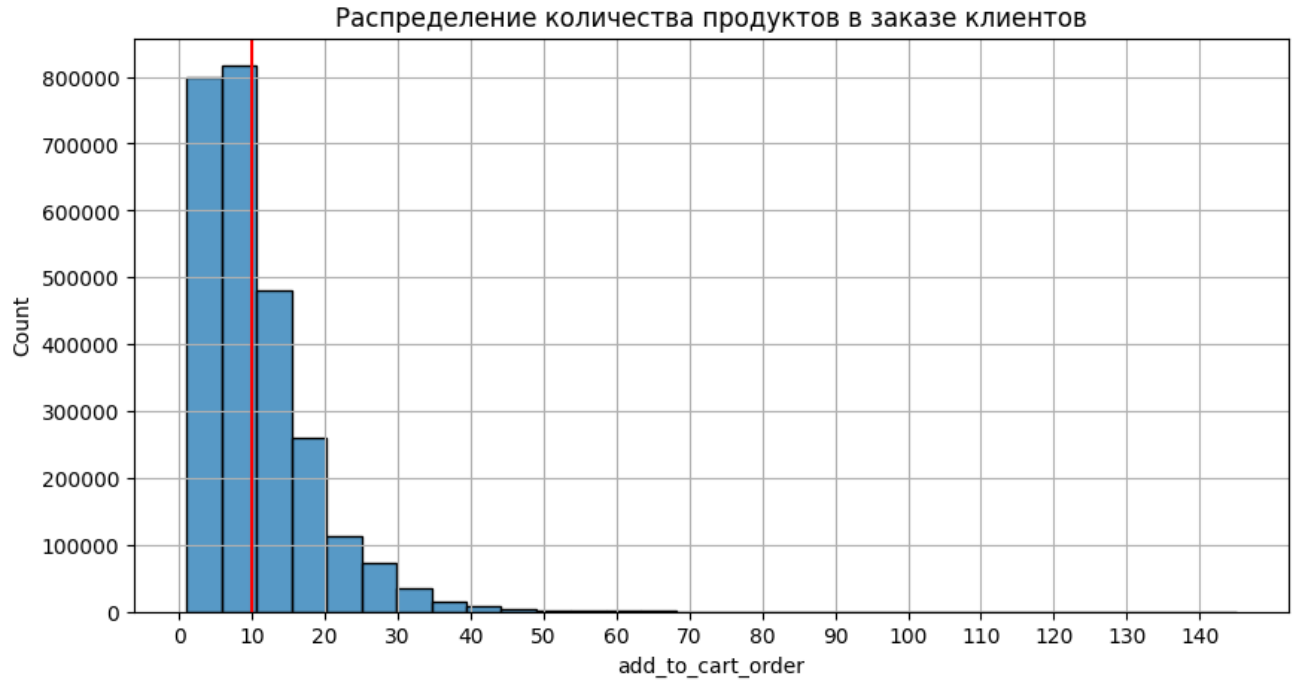
product_id : int - уникальный идентификатор товара
product_name: str - название товара
aisle_id : int - уникальный идентификатор подкатегории
department_id : int - уникальный идентификатор категории
aisle: str - название подкатегории
department : str - название категории

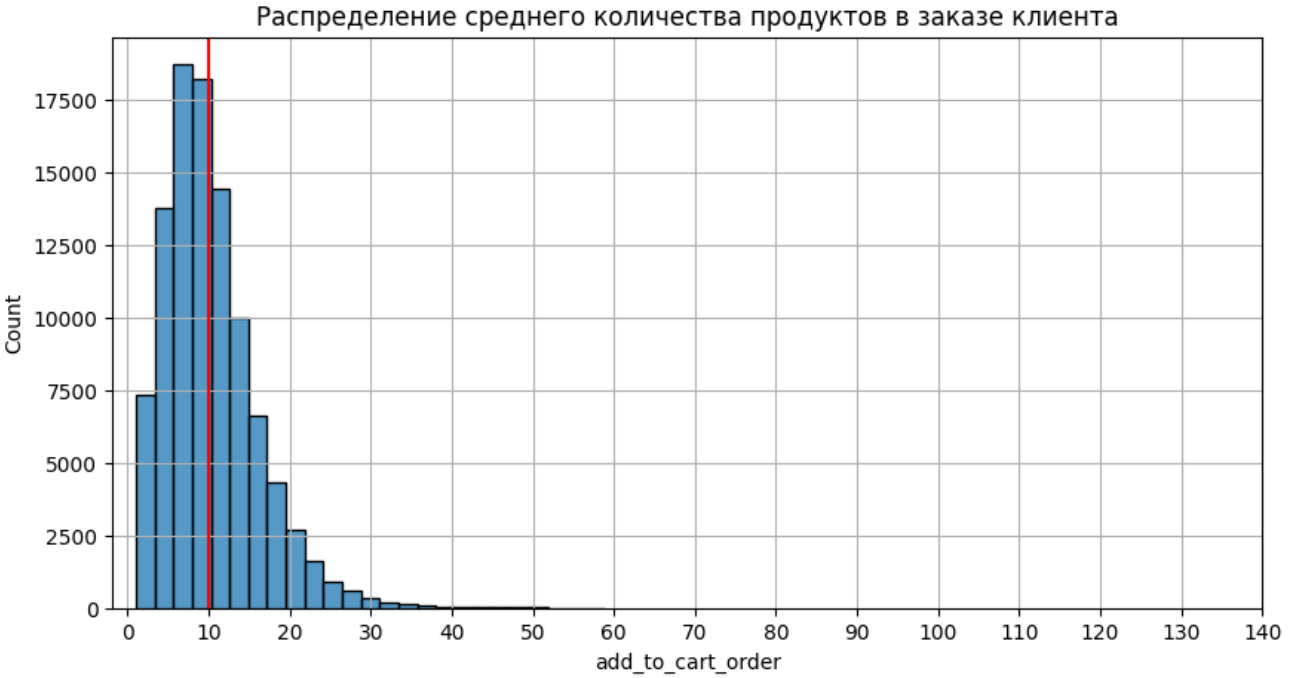
transactions.csv (26 408 073 строк, 100 000 уникальных пользователей):

order_id : int - уникальный идентификатор транзакции
user_id : int - уникальный идентификатор покупателя
order_number : int - номер транзакции в истории покупок данного пользователя
order_dow : int - день недели транзакции
order_hour_of_day : int - час совершения транзакции
days_since_prior_order : float (1045204 None) - количество дней с совершения предыдущей транзакции данным пользователем
product_id : int - уникальный идентификатор товара
add_to_cart_order : float - номер под которым данный товар был добавлен в корзину
reordered : float - был ли товар "перезаказан"

EDA

Из распределений количества товаров в корзине клиентов видно, что в среднем клиенты заказывают 10 товаров.
Далее рекомендательная система будет оцениваться с k=10.





ПОДГОТОВКА ДАННЫХ

Формируется разреженная матрицы интеракций по взаимодействию user-item

Для обучения классических ML моделей применялась матрица интеракций, где на пересечении строк с пользователями и столбцов с продуктами стоит бинарный признак, который соответствует наличию взаимодействия.
Пользователи и продукты кодируются по порядку следования идентификаторов.
Для оптимизации вычислительных ресурсов при работе с большим объёмом разреженных матриц применяются scipy sparse matrix.

```
1  X
```

✓ 0.0s

```
<99999x29454 sparse matrix of type '<class 'numpy.float64'>'
        with 9000952 stored elements in Compressed Sparse Column format>
```

```
1  X[90:100, 80:90].toarray()
```

✓ 0.0s

```
array([[0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 1., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])
```

COLLABORATIVE FILTERING

Написан класс для применения коллаборативной фильтрации (user based подход), где в качестве метрики схожести применялась корреляция Пирсона.

В рамках обучения collaborative filtering входными данными является подготовленная матрица интеракций. В качестве метрики схожести взята корреляция Пирсона.

$$s(u,v) = \frac{\sum_{i \in I_u \cap I_v} r_{ui} r_{vi}}{\sqrt{\sum_{i \in I_u} r_{ui}^2} \sqrt{\sum_{i \in I_v} r_{vi}^2}}$$

I_u – множество продуктов, купленных пользователем.

r_{ui} – факт приобретения продукта пользователем (0 или 1).

LIGHTFM И EASE

Альтернативные модели :

1) lightfm;
2) EASE.

С целью улучшения метрик качества было применено 2 альтернативных подхода:

1. Обучение модели из готового пакета lightfm.
2. Обучение модели EASE, реализовав алгоритм в собственном классе.

Алгоритм EASE заключается в приближении матрицы интеракций матрицей B:

$$\min_B \| X - XB \|_F^2 + \lambda \cdot \| B \|_F^2$$

$$diag(B) = 0$$

где B - матрицы (item x item).

РЕЗУЛЬТАТЫ

В ходе работы было реализовано 4 различных подхода к построению рекомендательных систем:

- Baseline Pop-based (топ из продуктов)
- Collaborative filtering
- LightFM
- EASE

Оценка моделей с k=10.

	Hitrate @k	AVG_Precision @k	AVG_nDSG @k	MNAP @k	MAP @k
Baseline Pop-based	0.438	0.062	0.015	0.008	0.006
Collaborative filtering	0.505	0.074	0.017	0.013	0.007
LightFM	0.450	0.067	0.018	0.013	0.008
EASE	0.632	0.110	0.018	0.012	0.007

В результате экспериментов предпочтение было отдано EASE.

Преимущества EASE модели:

1. Имеет решение в явном виде, нет необходимости учить итеративно;
2. Имеет всего 1 гиперпараметр и устойчив к его варьированию;
3. Теплый старт. Можно применять при появлении новой информации об интеракции пользователем без дообучения.

Недостатком считается, что при большом объеме продукции матрица B будет «тяжелой» и обучение потребует больших вычислительных ресурсов.

ПЛАНЫ

Разработка рекомендательной системы с применением нейросетевого подхода.

В рамках поэтапной разработки рекомендательной системы планируется применение нейросетевых подходов.
Логическим завершением этапа обучения ML моделей является последующие эксперименты в DL с применением матрицы полученной из алгоритма EASE в качестве входного embeding для продуктов.