Чтобы определить, как можно решить поставленную задачу, первым делом необходимо внимательно посмотреть на данные:

Изображение выглядит как текст, электроника, дисплей, снимок экрана

Автоматически созданное описание

Как видно из рисунка выше, более половины клиентов не идентифицировались, то есть, исходя из условий задачи, они не являются участниками клиентской программы. Это значит, что использовать подход с коллаборативной фильтрацией пользователей и искать похожих клиентов на всех данных не получится. Сразу же возникает проблема холодного старта. Обычно в таких случаях рекомендуют самые популярные товары (или другой объект, который нужно порекомендовать). Но в рамках данной задачи старт не совсем холодный, а скорее «тёплый», поскольку для каждой покупки имеется информация о других товарах в чеке, а угадать нам нужно дополнение к ним.

В данных есть половина известных пользователей, но в дальнейшем мы это все-таки проигнорируем. Хотя можно было бы скомбинировать модель из двух подходов: постоянным клиентам рекомендовать товары исходя из интересов похожих на них пользователей, а неизвестным клиентам просто предлагать популярные товары, каким-то образом похожие или связанные с теми, что есть у него в чеке.

Другим нюансом данной задачи является иерархия данных. Товары не просто существуют сами по себе, а сгруппированы по типу. Кроме того, стоит задача рекомендовать товары только из определенных групп.

Для решения этой задачи были разработаны некоторые эвристики, основанные на популярности товаров. В данном случае, как мне кажется, это оптимальный подход, поскольку объем данных довольно большой и сложные модели могут вычисляться на них слишком долго. К тому же нет гарантии, что сложные модели дадут хороший результат, поскольку половина пользователей неизвестна, а хороших признаков, которые можно было бы использовать в моделях, практически нет – есть пропуски по характеристикам товаров (в ноутбуке с решением на это можно посмотреть поподробнее). Иными словами, есть подозрение, что все модели так или иначе сведутся к кластеризации по группам, которые известны и так и к рекомендациям по популярности, которые можно вычислить проще.

После обработки данных было принято решение применить следующий подход:

1. Ищем ближайших соседей для всех групп товаров по признакам, которые мы отобрали и создали (подробнее в файле с решением).

2. Определяем группу, из которой покупатель совершил покупку товара.

3. Если эта группа та же, из которой нам нужно порекомендовать товары, рекомендуем 20 самых популярных товаров группы. (Во всех группах намного больше 20 товаров, поэтому будем считать, что они не закончатся).

4. Если есть несколько подходящих групп, откуда уже совершена покупка, рекомендуем 20 самых популярных товаров из самой популярной группы.

5. Если в чеке нет групп, из которых нам требуется рекомендовать товар, то берем ближайшего подходящего соседа-группу. Если куплено несколько товаров из разных групп, то берем соседа той группы, у для которой подходящий сосед ближе. Если у 2 х групп ближайший сосед стоит на одной позиции, то выбираем группу-соседа произвольно. Рекомендуем 20 самых популярных товаров из найденной группы.

Таким образом предложенное решение – это модифицированная версия рекомендации самых популярных товаров. Однако есть основания полагать, что такая система сработает немного лучше, чем классический popularity-based подход, поскольку она учитывает связь между группами товаров.

Задача достаточно интересная, ранее с рекомендательными системами так близко я не сталкивался, пришлось изучить довольно много теории. Если бы у меня было больше времени, я бы все-таки создал комбинированную систему, которая учитывает тот факт, что половина пользователей идентифицируемы. Тогда высока вероятность того, что рекомендации будут получше. К тому же расчет холодного старта можно делать на основе всех пользователей, поскольку он основан только на товарах, и покупках (событиях).

Если бы было меньше пропусков в номенклатуре, можно было бы применить какие-то NLP подходы, но, кажется, что в этом нет особого смысла, товары и так сгруппированы, а улавливать закономерности в названиях нецелесообразно в рамках данной задачи.

Если бы не было пропусков по id пользователей, наилучший результат скорее всего дала бы коллаборативная фильтрация по пользователям. Но с учетом того, что есть строгое ограничение по группам утверждать это точно нельзя.