

1. Данные.

Данные были отсортированы с целью найти какие-нибудь факторы, которые влияют на покупку той или иной категории, как например, день недели, выходной ли день, магазин, количество покупаемого топлива (больше, меньше среднего). Эти факторы рассматривались по частоте покупок категорий товаров. Как оказалось, тенденция почти всегда одна и та же

Кофейные напитки с молоком

Вода

Сладкие Уранированные напитки, холодный чай

Энергетические напитки

Снеки

Соки и сокосодержащие напитки

Кроме понедельника, когда немного выше стоят энергетические напитки.

Значит, имеет большее значение именно совместная встречаемость товаров.

При этом были рассмотрены и чаще совместно встречающиеся категории, и учет зарегистрированных покупателей. Оказалось, видимо, лучше делать просто по sku\_id (очевидно, статистика по товарам в каждом чеке является хорошим индикатором покупательского поведения).

1. Модели

Были рассмотрены возможности применения таких моделей

-сингулярное разложение матриц (по таблице частоты встречаемости всех товаров с нужными нам товарами, при этом эти частоты служили как бы «оценкой рейтинга»), что позволяет выбрать наиболее близкие между собой товары.

- полносвязная сеть Keras с эмбеддингами – нужных товаров, всех товаров – не получилась

-KMeans, который способен кластеризовать похожие чеки и на этой основе предлагать товары, но здесь оказалось сложно подобрать количество кластеров (мотод «локтя» дал какую-то гребенку).

- модель, как языковая. Для которой значениями была бы последовательность с 20 нужными товарами, а ключами – всевозможные сочетания остальных товаров в чеке. Но чеков с 20 нужными товарами оказалось очень мало – всего около 200.

- В итоге была выбрана частотная модель – просто частота встречаемости с нужными товарами, которая была дополнена коррелирующими товарами из сингулярного разложения, - если такие категории были – они ставились на первые места.

3.Если бы было больше данных

по пользователям – их материальному достатку – например, судя по авто что легко можно добыть на азс, возрасту и образу жизни - можно бы было лучше применить некоторые алгоритмы, легче определиться с количеством классов покупателей, выявить целевую аудиторию разных товаров. А также больше оперативной памяти – можно бы было посмотреть на сингулярное разложение матриц, группируя товары по чекам, например.