**Краткий отчет**

1. Модель рекомендации

В качестве финальной модели, выбрал модель основанную на рекомендацях популярных связанных товаров к текущему чеку. На основании предоставленных транзакций была подготовлена матрица взаимосвязей товаров (товары х товары) таким образом, что в столбцах были товары из чека, а в строках товары, которые были куплены с ним. К примеру, в чеке было топливо, вода и сигареты, в столбец матрицы заносится топливо а в строках сигареты и вода, а в значениях количество купленных позиций каждого товара в чеке, точно так же с водой и сигаретами, далее повторяется для всех чеков. В итоге получается матрица, которая работает следующим образом, приходит клиент, у него в чеке какой то товар или набор товаров, для каждого из товара находится соответствующий столбец матрицы (sku\_id), а в строках указаны другие товары, которые покупали с топливом и указано сколько раз такой товар покупали (информация со всех чеков), столбец сортируется по убыванию — это и есть рекомендации для товара. Точно так же формируются таблицы рекомендаций по каждому товару в чеке. Затем все эти таблицы рекомендаций складываются, рейтинг одинаковых товаров к рекомендации суммируется и получается топ по чеку, откуда берется 20 верхних товаров. Если в чеке товар, которого нет в матрице рекомендаций или товар редкий и у него не достаточно рекомендаций в матрице — то таким клиентам в качестве рекомендаций выдается топ по всей имеющейся матрице. Так же, за счет того, что один из самых популярных товаров, после топлива, это кофе, то рекомендательная система в качестве первых 5-8 рекомендаций выдает разные виды кофе, а с точки зрения жизни и логики, если клиент не хочет кофе, то нет смысла предлагать ему другие виды этого напитка, нужно предложить товар из следующей популярной категории, например вода, эту функцию я реализовал в финальной модели, формируется топ товаров, а затем по кругу предлагается по одному самому популярному товару из каждой категории, пока не наберется 20 рекомендаций, но на общей оценке она показала немного худший результат, чем рекомендовать «сначала 8 видов кофе», потому что это самый популярный товар и с точки зрения метрики, выгодней перебрать все варианты кофе и угадать тот что был в чеке, чем рекомендовать на второй позиции другую категорию (например воду). Так же, в среднем 10% чеков содержат более одной позиции товаров одной категории (из рекомендуемых), например две воды в чеке или два кофе, в жизни, для рекомендательной системы такие повторы нужно убирать, например если у человека в чеке уже есть кофе, то не нужно рекомендовать ему другой вид кофе, а с точки зрения метрики выгодней делать так.

Для метрики выбрал модель с повторами товаров из одной категории, а для жизни считаю, что повторы нужно исключать. Рекомендовать просто топ товаров всем подряд, тоже не худший вариант.

2. Что бы я сделал имея больше ресурсов/времени/данных.

- По изучал бы больше различных методов, алгоритмов и подходов создания рекомендательных систем, т.к. опыта в построении рекомендаций мало, за 2 недели смог применить только базовые вещи.

- Если в чеке один-два товара, то такому клиенту, если рекомендовать просто топ товаров, можно угодить примерно в 1 из 10 случаев, считаю, что улучшить показатель можно если формировать профиль клиента по карте лояльности (идентификатору), фиксируя предпочтения и вкусы клиента и предлагая его «любимые» товары. Так же не у всех клиентов есть ИД, в этом случае можно либо рекомендовать топ либо искать дополнительные признаки.

- Проработал бы различные варианты кластеризации. Можно клиентов с ИД кластеризовать по истории а новых клиентов пытаться отнести к какому то кластеру (например по чеку, какие и сколько позиций, какой и сколько вида топлива покупает, какой у него автомобиль и др), для новых клиентов рекомендовать популярные товары из кластера, куда его отнесем. Возможно стоит текущие категории немного укрупнить как доп признак.

- Провел бы более детальный анализ, где ошибается моя рекомендательная система и что можно «подкрутить». Проверил бы гипотезы о количестве заправляемых литров и вида топлива и товаров в чеке, поигрался бы с весом категорий товаров или отдельных товаров.

- Попробовал ансамбли из моделей, можно нагенерить признаков, типа количество заправок, сезонность продажи товаров, праздники, дни недели и поискал бы взаимосвязи, если бы было время чеков, можно было бы поискать взаимосвязи от времени суток (утром вероятно кофе продается лучше, а вечером вода или энергетик), погодные условия и температура, результаты рекомендаций от других моделей, все это «запихнуть» в какой-нибудь xgboost или нейросеть (у меня тут мало опыта, с удовольствием бы по изучал).

- Провел бы анализ в разрезе заправок, возможно на каких то заправках совсем другой топ товаров и стоит предлагать его, вместо общих рекомендаций.

- Интересное наблюдение, 33% чеков без топлива, достаточно много людей заезжают на заправку для покупки других товаров, топ из них это кондитерские изделия, сезонные товары и автотовары, перекус, не малый процент от продаж, возможно стоит разместить такие товары на тех заправках где их нет, по исследовал бы взаимосвязи тут.

3. Данные.

- В данных, которые были предоставлены для анализа и подготовки системы рекомендации обнаружил несколько недочетов, помимо отсутствующих значений (их не так много), большое количество записей имеет значение «unknown», что может говорить о недостаточном качестве данных, стоит обратить на это внимание и устранить, т.к. эти данные вполне могут быть использованы для улучшения модели. Так же у достаточного количества товаров (транзакций) не являющихся топливом указаны литры, а их в этом случае не должно быть, это может намекать в целом на сомнения корректности заполнения столбца «litrs».