Пояснение к **Проекту по курсу ПММО на тему: «Uplift-моделирование + Фильтр Калмана»**

Хубежова Ольга,
Андросова Никита,
Третьяк Диана

Бизнес-постановка

Дано: Существует два города: богатый (Y) и бедный (X). По результатам проведения рекламной кампании для богатого города была оценена uplift-модель.

Цель: понять, какие результаты будут после проведения рекламной кампании в бедном городе. Для бедного города есть только информация о сумме чеков обезличенных пользователей за определенный период.

Задачи:

1. подготовка необходимых данных, их предобработка
2. восстановление скрытых характеристик покупателей в бедном городе X
3. оценка uplift-модели для бедного города X

Фактический ход работы

1. подготовка необходимых данных, их предобработка

Взят датасет x5/retail_hero, из которого мы сформировали богатый и бедный город. Человек считался жителем богатого города, если общая сумма его покупок была

больше или равна медиане по всем потребителям.

Также жители богатого города были случайно разделены на контрольную и целевую группу. В используемом датасете представлены данные по пользователям до распространения коммуникации, поэтому для выполнения предпосылок авторов к таргетному признаку целевой группы был прибавлен шум, экспоненциально зависящий от возраста покупателя.

Учитывая изначальную структуру датасета (чеки + количество товара в них + id товара), пришлось пересобрать структуру датасета и нагенирировать еще фичей, для большего понимания поведения клиента в течение дня и в рамках каждого чека. Нагенирировали около 30 фич, после чего был проведен анализ “полезности” фичи для нашей задачи - подходит ли фича для наших моделей, коррелирует ли она с другими, дает ли улучшение в модели.

Первоначальным фичам, как и по одобренным для дальнейшей работы, был проведен EDA, по одобренным - еще и препроцессинг. В рамках EDA выяснили, что у всех фичей невероятно длинные хвосты, присутствуют “странные” и нелогичные значения. Был проведен поиск дублей, отсутствующих значений, взаимных корреляций, визуальный анализ распределений, боксплоты. В рамках препроцессинга пробовали откидывать значения по IQR, но откинулось слишком много - сделали выбор в пользу винсзоризации на уровне 5% по обеим границам. Так же пробовали сделать трансформер для данных, однако решили отказаться от этой идеи в пользу интерпретируемости данных

2. восстановление скрытых характеристик покупателей в бедном городе X

Обучение модели пространства состояний

Известны некоторые данные по потребителям богатого города Y. Данные имеют панельную структуру, то есть для каждого потребителя известна информация о номере и времени транзакции, сумме и составе покупки.

Таким образом, каждого пользователя можно рассматривать как динамическую систему, которую можно описать моделью пространства состояний. Данные были усреднены по дням, и итого получены временные ряды за 118 дней. Была оценена модель пространства состояний. Для конкретного случая ее можно описать уравнениями:

$$x_{t+1} = Ax_t + w_t$$

A – матрица перехода состояний, wt– гауссовский шум состояния с нулевым средним и ковариацией Q,

$$y_t = Cx_t + v_t$$

C – матрица наблюдений, vt- гауссовский шум наблюдений с нулевым средним и ковариацией R.

В качестве состояния модели подается на вход матрица имеющихся признаков города Y, уравнение (1) оценивается МНК. Затем МНК оценивается уравнение наблюдения (2),

где наблюдениями выступали средние суммы чека за день. Шумы состояния и наблюдения оцениваются как остаток после применения матриц A и C соответственно.

$$w_t = x_{t+1} - Ax_t \qquad v_t = y_t - Cx_t$$

Таким образом, на выходе были получены оценки матриц A, C, Q, R, определяющие линейную модель пространства состояний на богатом городе.

Восстановление скрытых состояний фильтром Калмана
Оцененные матрицы использовались в дальнейшем при инициализации фильтра Калмана для восстановления пропущенных значений для потребителей бедного города по фактически наблюдаемым данным, в нашем случае, по информации о сумме чека за период.
В качестве начального значения для каждого прогнозируемого признака выбиралось среднее значение по каждому признаку на выборке богатого города, затем на первом шаге с помощью матриц A и Q предсказывалось последующее значение состояния и его ковариация.

На втором шаге предсказанное состояние и его ковариация корректировались с учетом полученного наблюдаемого значения yt (сумме чека в этот момент времени) и коэффициента Калмана. Описанная процедура происходила итеративно по доступному временному ряду для определенного потребителя. Так, предполагается, что с каждым шагом индивидуальные характеристики (скрытые состояния) потребителя восстанавливаются, учитывая динамику его покупок.

Так как оценивание модели пространства состояний происходило по усредненным за день характеристикам

богатых жителей, то есть смысл немного увеличить значения матрицы шума состояний Q, чтобы при корректировке прогнозов модель больше опиралась на наблюдаемые значения потребителя, а не на общую усредненную динамику, которую модель “выучила”. К матрице Q по главной диагонали был прибавлен шум 0,15. С помощью такой корректировки параметра модель получилась более адаптивной.

3. оценка uplift-модели для бедного города X

Как мы сказали после завершения первого этапа: восстановление характеристик (“скрытых” состояний) жителей города X (бедный город) с помощью LSS и Фильтра Калмана необходимо понять, как наш пилот проявит себя в этом городе.

Для этого необходимо оценить uplift модель на выборке где у нас был пилот и контроль - то есть “богатый” город. Как мы сказали ранее для жителей Y|T=1 был добавлен синтетический шум, связанный эксп. зависимостью с возрастом покупателя. Итого, нам необходимо для оценки uplift модели:

1) Yi - значение целевой метрики для клиента i. В качестве такой мы воспользовались биноризованной суммой покупок за день. Если purchase_sum_i > E(purchase_sum), то 1, иначе 0
2) Xi - характеристики пользователя (наши фичи которые мы собрали на этапе обработки данных)
3) Ti - флаг тритмента, то есть факт того, что человек в пилоте (например, мы с ним провзаимодействовали)

На этих данных мы воспользовались Transformed Target подходом, который модицифирует целевую переменную на факт попадания в пилот и характеристики пользователя. Таким образом мы обучили модель на богатом городе в

надежде уловить, то как совокупность состояний пользователя и факт инициативы (T) влияют на целевой показатель.
Далее обученную модель мы понесли в “бедный” город и оценили на сэмпле из клиентов (30к - например, в силу бизнес ограничений) uplift. Того итогом стало понимание, как инициатива по взаимодействию с клиентами сработает на городе о котором мы в силу обстоятельств почти ничего не знаем.