Введение в Uplift-моделирование

Елена Кантонистова

Коммуникация

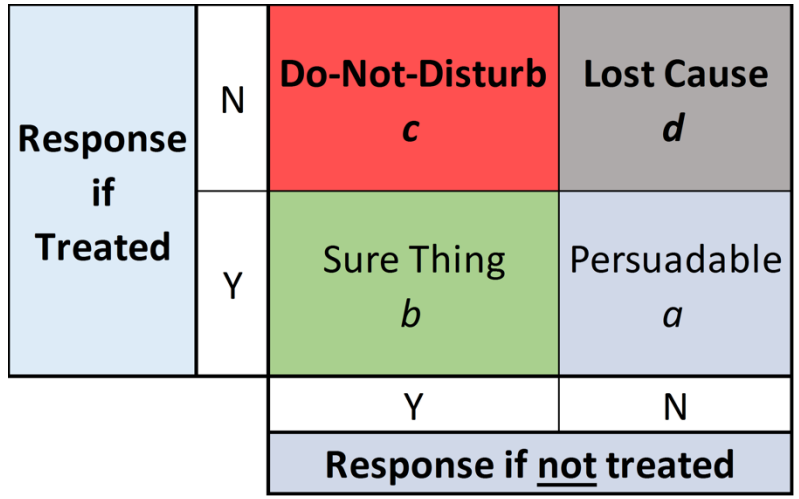
- Часто возникают задачи, связанные с определением клиентов, которым необходимо отправить ***коммуникацию*** с целью стимулировать совершить целевое действие (например, купить тот или иной продукт). Отправка коммуникации, во-первых сама по себе стоит определенных средств, а с другой стороны часть клиентов может реагировать на них негативно: до коммуникации клиент был готов купить товар, а после коммуникации такое желание у него пропадает.
- Далее под ***коммуникацией (или маркетинговой коммуникацией)*** может пониматься отправка SMS, Push-сообщений в мобильном приложении, email и пр.

Типы клиентов

Клиентов в части отношения к коммуникации можно разделить на 4 основных типа:

- **"Не беспокоить" (Do-Not-Disturbs)** - это те, клиенты, которые будут негативно реагировать на коммуникацию. То есть они с большей вероятностью купили бы товар, если бы коммуникация не была направлена. Коммуникация с этими людьми тратит маркетинговый бюджет (прямые затраты) и уменьшает продажи (косвенные убытки)
- **"Потерянные" (Lost Causes)** и **"Лояльные" (Sure Things)** - это те клиенты, которые не купят или купят товар независимо от наличия коммуникации. То есть коммуникация с данными людьми расходует маркетинговый бюджет
- **Убеждаемый (Persuadables)** - это клиенты, которые положительно реагируют на проведенную маркетинговую коммуникацию. То есть человек купит товар только при условии направления ему коммуникации.

Таким образом, наша задача найти именно Убеждаемых клиентов.



Как найти убеждаемых клиентов?

1. Давайте научим нашу модель машинного обучения искать тех клиентов, которые **похожи на тех**, кто купил товар после получения коммуникации. Но в этом случае мы **найдем и убеждаемых клиентов** (купили из-за получения коммуникации) **и лояльных клиентов** (в любом случае купили бы наш товар). Наша модель никак не сможет отличить убеждаемых клиентов от лояльных клиентов.

2. Можно для каждого клиента (клиента под номером i) прогнозировать реакцию на коммуникацию: $\tau_i = Y_i^1 - Y_i^0$, где Y_i^1 - потенциальная реакция человека, если бы с ним была бы коммуникация, а Y_i^0 - потенциальная реакция человека, если бы коммуникации не было. То есть обучить модель прогнозировать следующий таргет:

- $\tau_i = 0$, если поведение человека после получения коммуникации не изменилось
- $\tau_i = 1$, если человек купил товар после получения коммуникации, а без неё не купил бы
- $\tau_i = -1$, если бы человек после получения коммуникации передумал покупать товар

Это неудачные формулировки.

Как найти убеждаемых клиентов?

Мы будем прогнозировать **насколько изменится вероятность покупки** товара клиентом с заданным признаковым описанием Xi от того, что мы отправим клиенту коммуникацию.

Факт отправки/не отправки коммуникации будет отдельным признаком.

Чтобы вычислить эту вероятность, нужно:

- во-первых, спрогнозировать вероятность того, что данный клиент купит товар при наличии коммуникации $P(Y = 1|X = X_i, W = 1)$, где Y - это флаг того, что клиент купил товар, W - флаг того, направили ли ему коммуникацию
- во-вторых, спрогнозировать вероятность того, что данный клиент купит товар при отсутствии коммуникации $P(Y = 1|X = X_i, W = 0)$
- вычесть полученные результаты

То есть

$$uplift(X_i) = P(Y|X = X_i, W = 1) - P(Y|X = X_i, W = 0)$$

Практика-1

https://colab.research.google.com/drive/1b-r7L2w_aobRe1JM5QJRwxY_n1NdB7L3?usp=sharing

Метрики качества Uplift: Uplift@k

Uplift@k - одна из самых простых и понятных метрик. Отсортируем клиентов по размеру предсказанного Uplift, возьмем первые **k%** клиентов и вычислим разницу между средним значением таргета (значение целевой переменной y, также еще называют response_rate) в целевой и контрольной группе. Часто нам не нужно качество на всей выборке, а нас интересует только некоторый верхний процент клиентов.

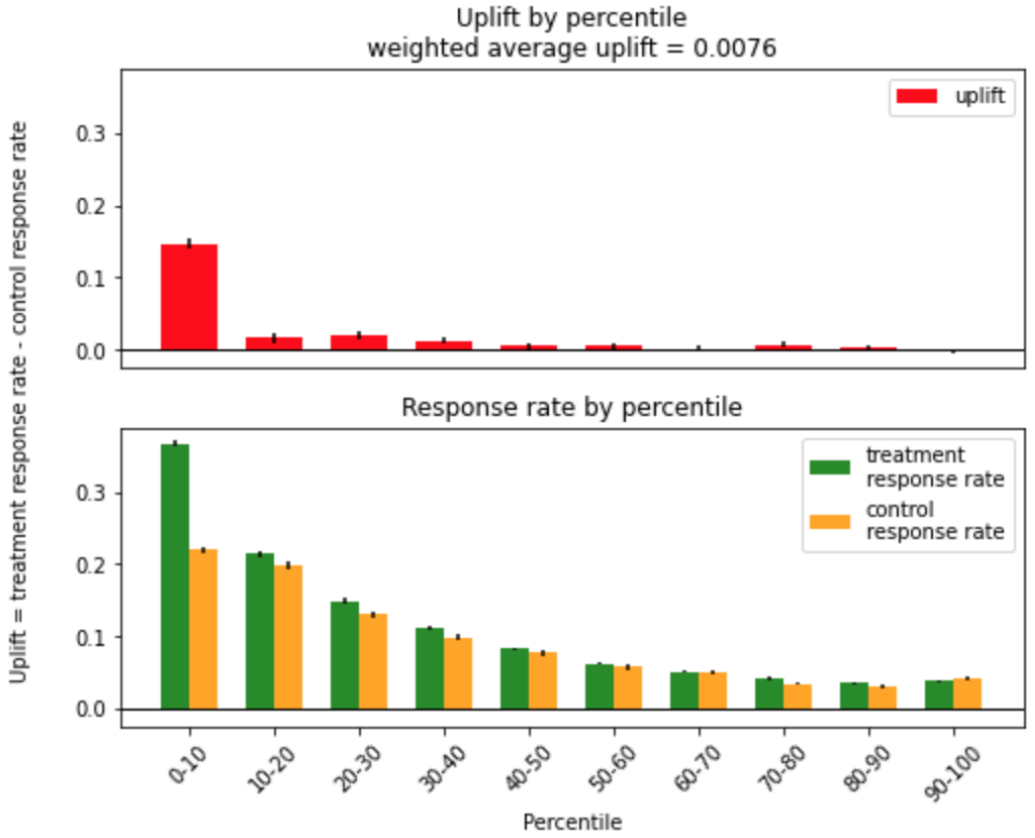
$$uplift@k = response\_rate@k_{w=1} - response\_rate@k_{w=0}$$

При формулировке этого определения есть некоторая двусмысленность:

1. Мы сначала сортируем все объекты по предсказанному Uplift. Затем выбираем **k%** лучших, отделяем контрольную и целевую группу, затем считаем среднее значение y по каждой из групп и вычитаем. В этом случае, соотношение контрольной/целевой группы может быть неравным. Например, пусть у нас 100 наблюдений и k=10, в верхние 10 значений по предсказанному uplift может попасть 6 значений из контрольной группы и 4 значения из целевой
2. Мы сначала отделяем контрольную и целевую группу, затем каждую из них сортируем по предсказанному Uplift, берем верхние k% в каждой группе, считаем средние и вычисляем Uplift.

Uplift by percentile

Естественным обобщением Uplift@k метрики является ***Uplift by percentile***. То есть мы сортируем объекты по предсказанному Uplift, делим на бины (то есть например [0-20, 20-40, 60-80, 80-100]) и для каждого бина вычисляем свой аналог Uplift@k. Эти значения можно визуализировать на графике (реализован готовый класс sklift.viz.plot_uplift_by_percentile).



Uplift curve

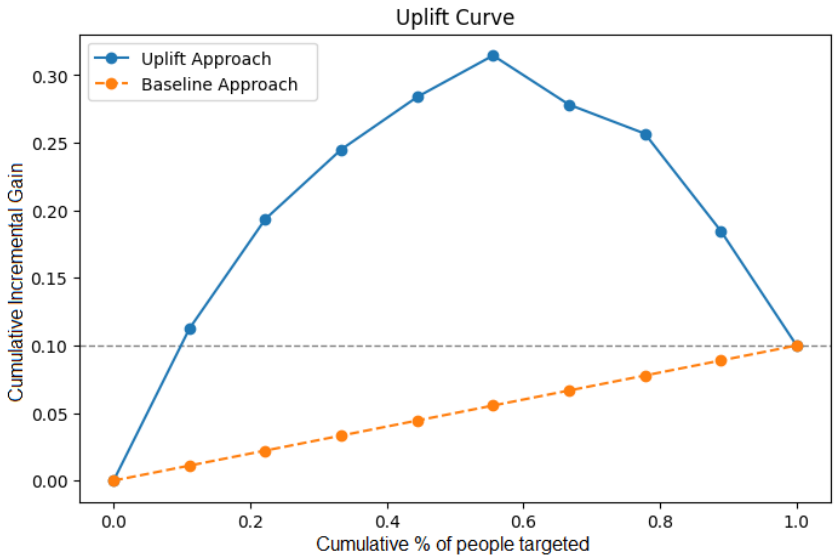
- **Сортируем всех пользователей по предсказанному uplift (от большего к меньшему)**. Это означает, что **самые "восприимчивые" к воздействию пользователи идут первыми**.
- **Разбиваем выборку на квантили**:
  - Первая квантиль (например, топ-10% пользователей с наибольшим uplift)
- Для каждой точки считаем uplift как разницу **конверсий между Treatment и Control**:

$$Uplift = \frac{\text{conversion rate в Treatment}}{\text{conversion rate в Control}} - 1$$

Где:

- **Conversion Rate (CR)** = $\frac{\text{количество конверсий}}{\text{общее число пользователей}}$.

  - Первая точка – первая квантиль (например, топ-10%)
  - Вторая точка – первые две квантили (топ-20%) и так далее
- Строим Uplift-кривую:
  - По оси **X** → Доли пользователей (кумулятивно).
  - По оси **Y** → Кумулятивный uplift (разница в конверсиях).



Чем выше кривая — **тем лучше модель выделяет аудиторию, на которую воздействие дает максимальный эффект**.

Qini curve

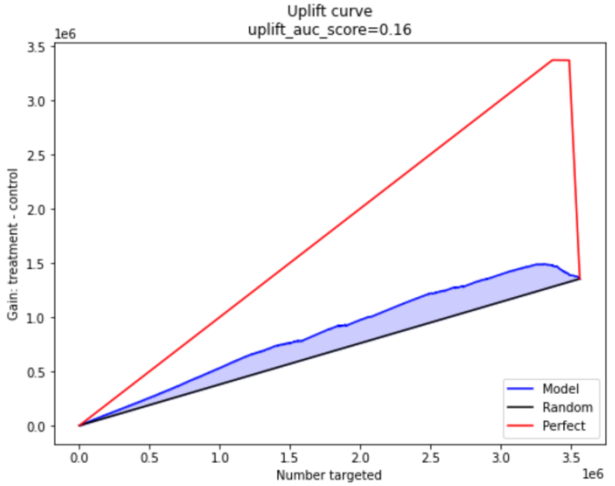
- **Qini Curve** (кривая Кини) — это улучшенный вариант **Uplift Curve**, который учитывает **разные размеры контрольной и тестовой групп**.
- **Разница с Uplift Curve** → Qini Curve корректирует uplift относительно размера групп, чтобы избежать искажений, если группы имеют **разное количество пользователей**.

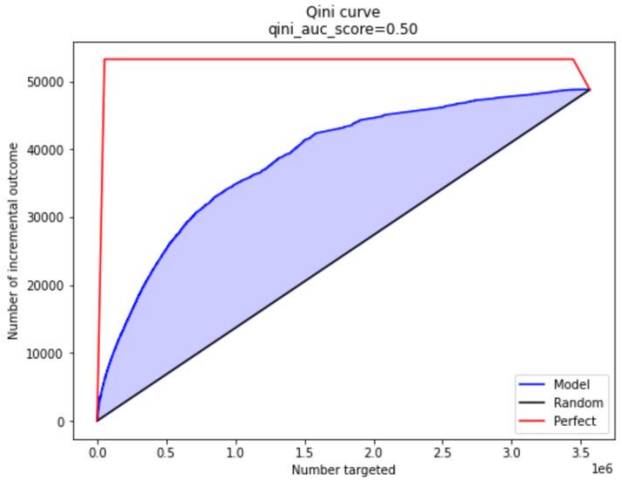
Алгоритм как в Uplift-curve, но Uplift будем считать с учетом размеров контрольной и тестовой групп:

$$\text{Qini Gain} = \sum_{i=1}^{k} (N_T^i \cdot CR_T^i - N_C^i \cdot CR_C^i)$$

Где:

- N_T^i — число пользователей в Treatment в i-м квантиле.
- CR_T^i — конверсия в Treatment в i-м квантиле.
- N_C^i — число пользователей в Control в i-м квантиле.
- CR_C^i — конверсия в Control в i-м квантиле.





Практика-2

https://colab.research.google.com/drive/1KZVXBykWBJBh7hzvDDDQ9fykIUnl1b3q?usp=sharing