

URLIFT-моделирование

Артем Савельев

Обо мне

- Меня зовут Артем Савельев
- Я работаю Data Scientist'ом в Сбербанке
- Наша команда отвечает за прикладной AI при продаже инвестиционных продуктов
- В прошлом году одной из моих задач была доработка существующего AutoML пайплайна под UPLIFT моделирование







План

1. Ресурсы
2. Uplift и классические задачи ML
3. Флаг коммуникации
4. Подходы к решению задач: metalearners
5. Практика 1
6. Подходы к решению задач: metalearners (продолжение)
7. Преобразование классов
8. Uplift-деревья
9. Метрики качества Uplift моделирования
10. Практика 2
11. Вопросы

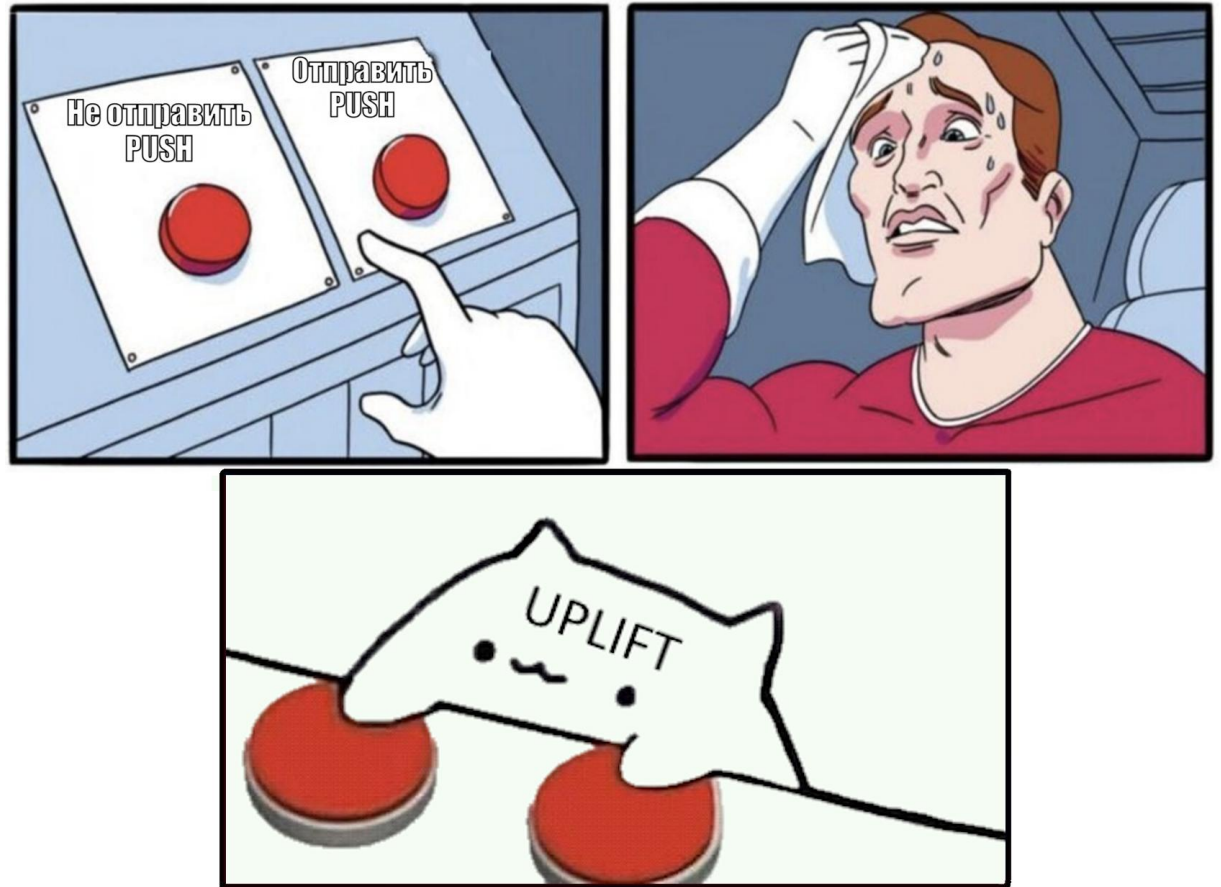
Recap

- Наша задача – оценить чистый эффект от коммуникации
- 4 типа клиентов по отношению к коммуникации
- Постановка задачи: $\text{uplift}(X_i) = P(Y|X=X_i, W=1) - P(Y|X=X_i, W=0)$
- Метрики качества: $\text{uplift}@k$, Uplift by percentile, qini, uplift curve

RENEW IF TREATED	Yes	Persuadables 	Sure Things 
	No	Lost Causes 	Sleeping Dogs Zzz 
		No	Yes
RENEW IF NOT TREATED			

Uplift и классические задачи

- Модель склонности к покупке скорее всего будет искать лояльных клиентов
- Прогноз реакции на коммуникацию невозможен так одному и тому же человеку нельзя и отправить, и не отправить коммуникацию
- В отличие от классических задач **нет правильных ответов**
- Нет возможности оценить качество модели привычным образом
- Новая сущность – флаг коммуникации, W , помимо признаков и таргета



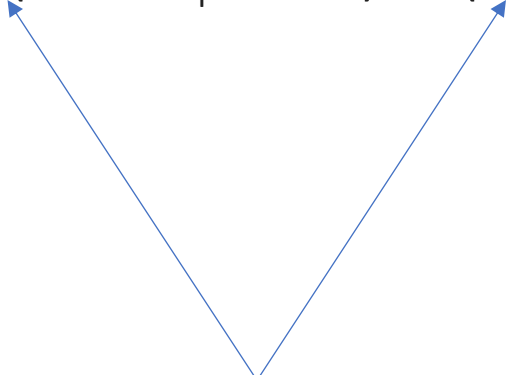
Флаг коммуникации

- $W=1$ — целевая группа, $W=0$ — контрольная группа (те люди, которые не получали коммуникацию)
- И контрольная, и целевая группа должны входить в тренировочный и тестовый датасет
- Клиенты из контрольной и целевой группы должны быть однородны, то есть быть из одной популяции, разбиение не должно зависеть от признаков
- Например, если вы продаете, игровые приставки и у вас в целевой группе люди возраста 20-30 лет из Москвы, а в контрольной — пенсионеры из сельской местности, то эффект кампании будет завышен и Uplift-модель на этом строить нельзя
- На практике иногда контрольная и целевая группа разных кампаний по одному продукту могут пересекаться и на таких данных нельзя строить модели

Подходы к решению задачи: Metalearners

Подход заключается в том, что мы прогнозируем каждую из вероятностей в формуле для Uplift с помощью классических ML-моделей, а затем вычисляем Uplift.

$$\text{uplift}(X_i) = P(Y|X=X_i, W=1) - P(Y|X=X_i, W=0)$$



Solomodel – обучаем одну модель и каждого клиента предиктим дважды с коммуникацией и без неё. Флаг коммуникации – признак, на котором обучаемся, **но для предикта мы его не используем**

Подходы к решению задачи: Metalearners

Подход заключается в том, что мы прогнозируем каждую из вероятностей в формуле для Uplift с помощью классических ML-моделей, а затем вычисляем Uplift.

The training process:

$$\text{fit} \left(\begin{array}{ccc|c|c} x_{11} & \cdots & x_{1k} & w_1 & y_1 \\ \vdots & \ddots & \vdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nk} & w_n & y_n \end{array} \right)$$

$X_{train} \quad W_{train} \quad Y_{train}$

The process of applying the model:

$$\text{predict} \left(\begin{array}{ccc|c} x_{11} & \cdots & x_{1k} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{m1} & \cdots & x_{mk} & 1 \end{array} \right) - \text{predict} \left(\begin{array}{ccc|c} x_{11} & \cdots & x_{1k} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ x_{m1} & \cdots & x_{mk} & 0 \end{array} \right) = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$$

$X_{test} \quad W_1 \quad X_{test} \quad W_0 \quad \text{uplift}$

Практика – построение первой Uplift модели

Подходы к решению задачи: Metalearners

Подход заключается в том, что мы прогнозируем каждую из вероятностей в формуле для Uplift с помощью классических ML-моделей, а затем вычисляем Uplift.

$$\text{uplift}(X_i) = P(Y|X=X_i, W=1) - P(Y|X=X_i, W=0)$$

Model_t

Model_c

Twomodel – обучаем две
независимые/зависимые модели на
целевой и контрольной группе

Подходы к решению задачи: Metalearners

Подход заключается в том, что мы прогнозируем каждую из вероятностей в формуле для Uplift с помощью классических ML-моделей, а затем вычисляем Uplift.

The training process:

$$\begin{array}{c} model^T = fit \left(\begin{array}{ccc} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{array}, \begin{array}{c} y_1 \\ \cdots \\ y_p \end{array} \right), \quad model^C = fit \left(\begin{array}{ccc} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{q1} & \cdots & x_{qk} \end{array}, \begin{array}{c} y_1 \\ \cdots \\ y_q \end{array} \right) \\ X_{train_treat} \quad Y_{train_treat} \qquad \qquad X_{train_control} \quad Y_{train_control} \end{array}$$

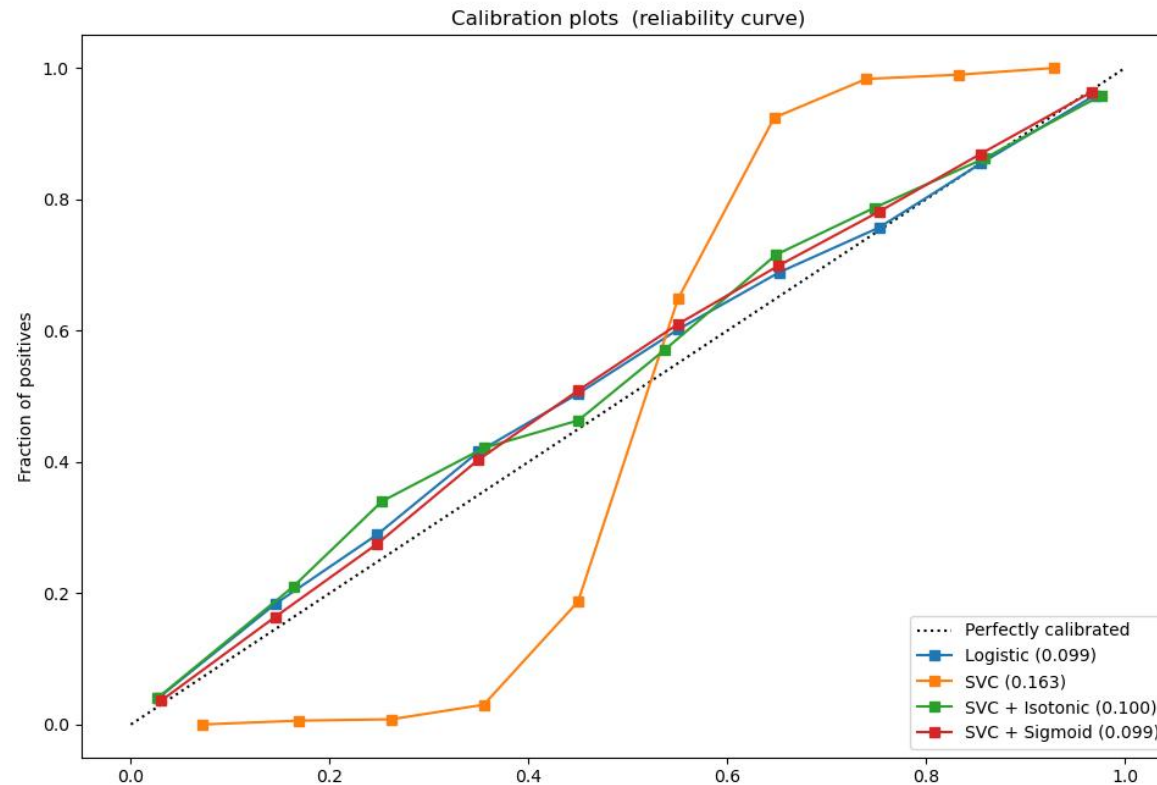
The process of applying the model:

$$\begin{array}{c} model^T \\ predict \\ proba \end{array} \left(\begin{array}{ccc} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{array} \right) - \begin{array}{c} model^C \\ predict \\ proba \end{array} \left(\begin{array}{ccc} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{array} \right) = \begin{array}{c} u_1 \\ \vdots \\ u_m \end{array}$$

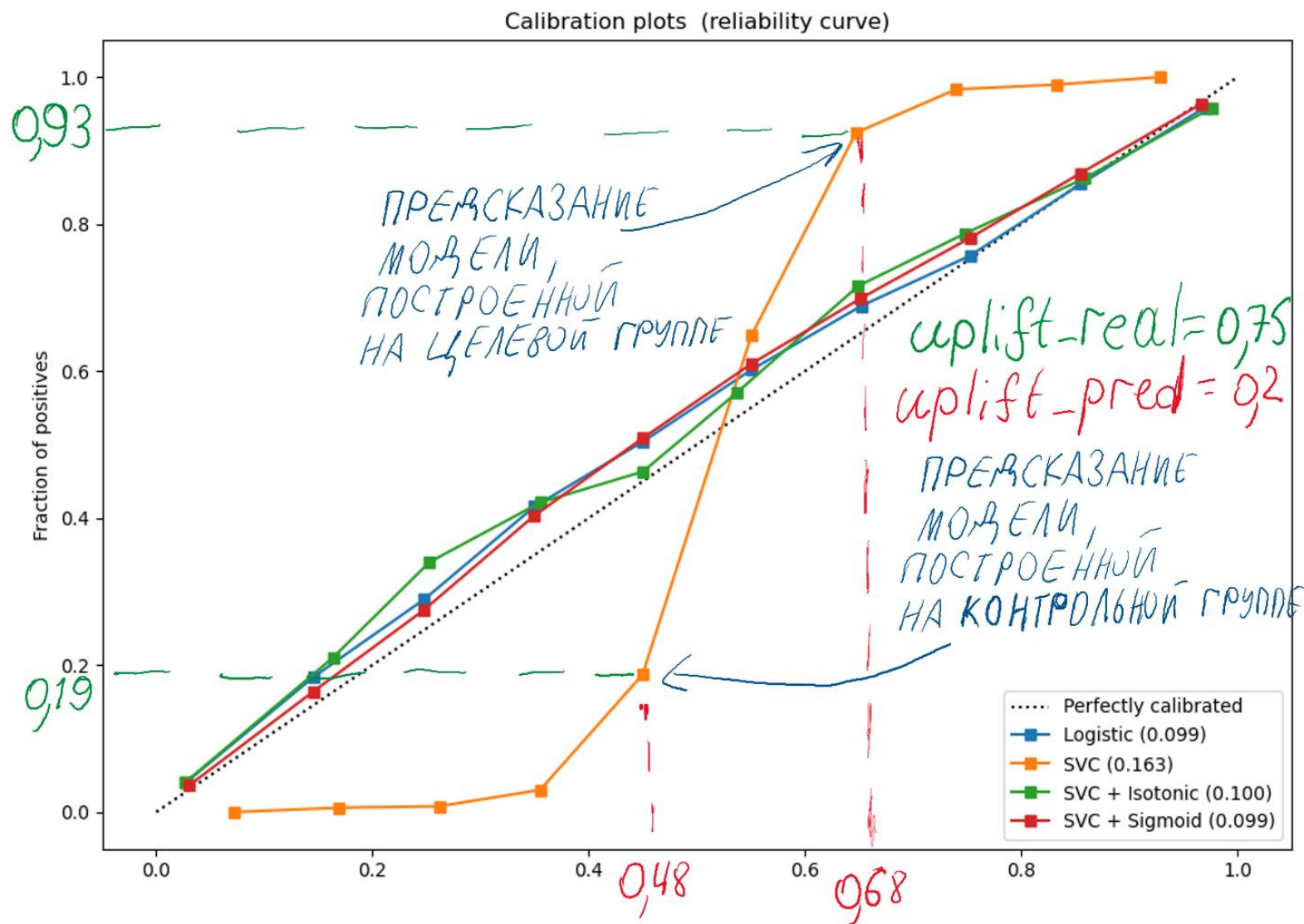
$X_{test} \qquad \qquad X_{test} \qquad \qquad uplift$

Подходы к решению задачи: Metalearners

- Для Uplift-моделей критически важна калибровка моделей.
- Все ML модели кроме логистической регрессии не выдают истинных вероятностей



Подходы к решению задачи: Metalearners



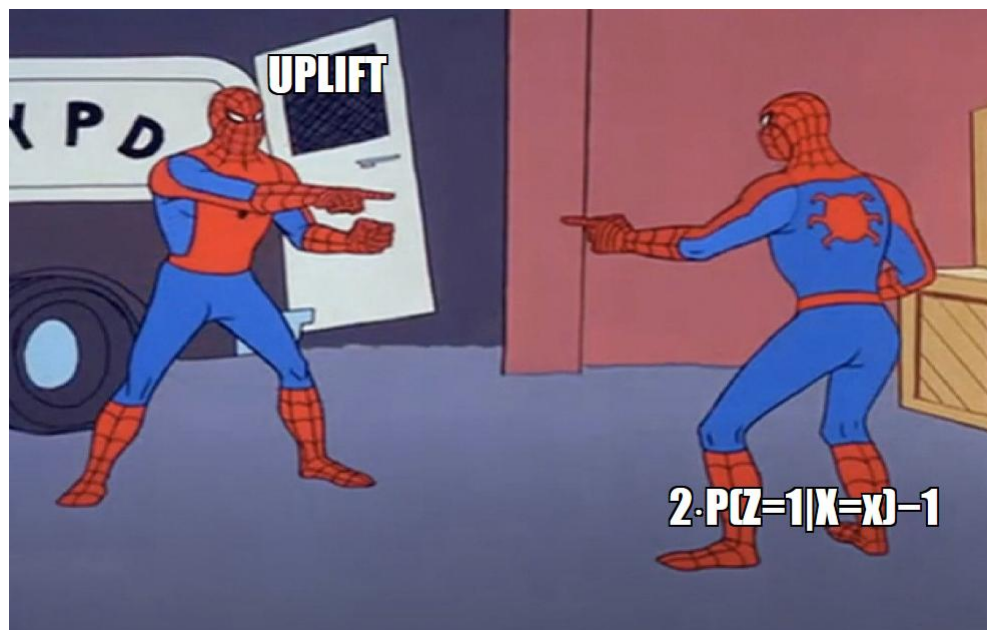
Подходы к решению задачи: трансформация классов

Подход заключается в замене переменной для прогнозирования, а именно

$$Z_i = Y_i * W_i + (1 - Y_i) * (1 - W_i)$$

Новая переменная равна 1, когда $Y=W$ и равна нулю, когда Y не равен W .

Покажем, что при определенных условиях построение прогноза $P(Z=1 | X=x)$ для Z тождественно прогнозированию $\text{uplift}(X_i) = P(Y|X=x, W=1) - P(Y|X=x, W=0)$



Подходы к решению задачи: трансформация классов

Распишем, что такое $P(Z=1 | X=x)$

$$P(Z=1 | X=x) = P(Z=1 | X=x, W=1) * P(W=1 | X=x) + P(Z=1 | X=x, W=0) * P(W=0 | X=x) =$$

$$P(Y=1 | X=x, W=1) * P(W=1 | X=x) + P(Y=0 | X=x, W=0) * P(W=0 | X=x)$$

Отметим, что $P(W=1 | X=x) = P(W=1)$ и $P(W=0 | X=x) = P(W=0)$, так как разбиение на контрольную и целевую группу не зависит от значений признака. Кроме того, если размер контрольной и целевой группы одинаков, то $P(W=1) = P(W=0) = 1/2$.

Таким образом,

$$P(Z=1 | X=x) = \frac{1}{2} * P^T(Y=1 | X=x) + \frac{1}{2} * P^C(Y=0 | X=x)$$

$$2 * P(Z=1 | X=x) = P^T(Y=1 | X=x) + 1 - P^C(Y=1 | X=x) = uplift + 1$$

Следовательно:

$$uplift = 2 * P(Z=1 | X=x) - 1$$

Подходы к решению задачи: трансформация классов

Обобщение на тот случай, когда размер контрольной и целевой групп не одинаков, а переменная Y_i не является бинарной, выглядит следующим образом:

$$Z_i = Y_i * (W_i - p) / (p * (1 - p))$$

где p – это вероятность отнесения к целевой группе:

$$p = P(W_i = 1 | X_i = x)$$

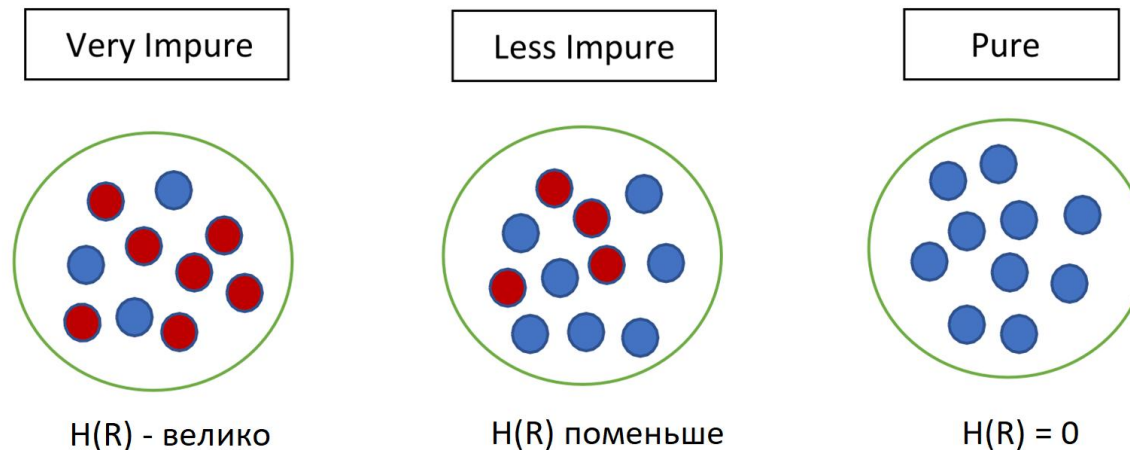
Эту величину можно оценить, как долю объектов с $W=1$ в выборке. Или обучить модель на X, W , которая будет прогнозировать p для каждого объекта. После применения формулы получаем новую целевую переменную, можем обучить модель регрессии с функционалом ошибки MSE.

Подходы к решению задачи: Uplift-деревья

Вспомним, как обучалось обычное решающее дерево

1. Выбираем критерий информативности $H(R)$, например энтропию, который некоторым образом характеризует однородность выборки
2. Задаём критерий остановки алгоритма
3. Жадно перебираем всевозможные признаки и пороги, чтобы максимизировать прирост информации (Information Gain). Рекурсивно повторяем процедуру до того момента, как дойдем до точки останова.

$$IG = H(R) - |R_l/R|(H(R_l) - |R_r/R| * H(R_r))$$

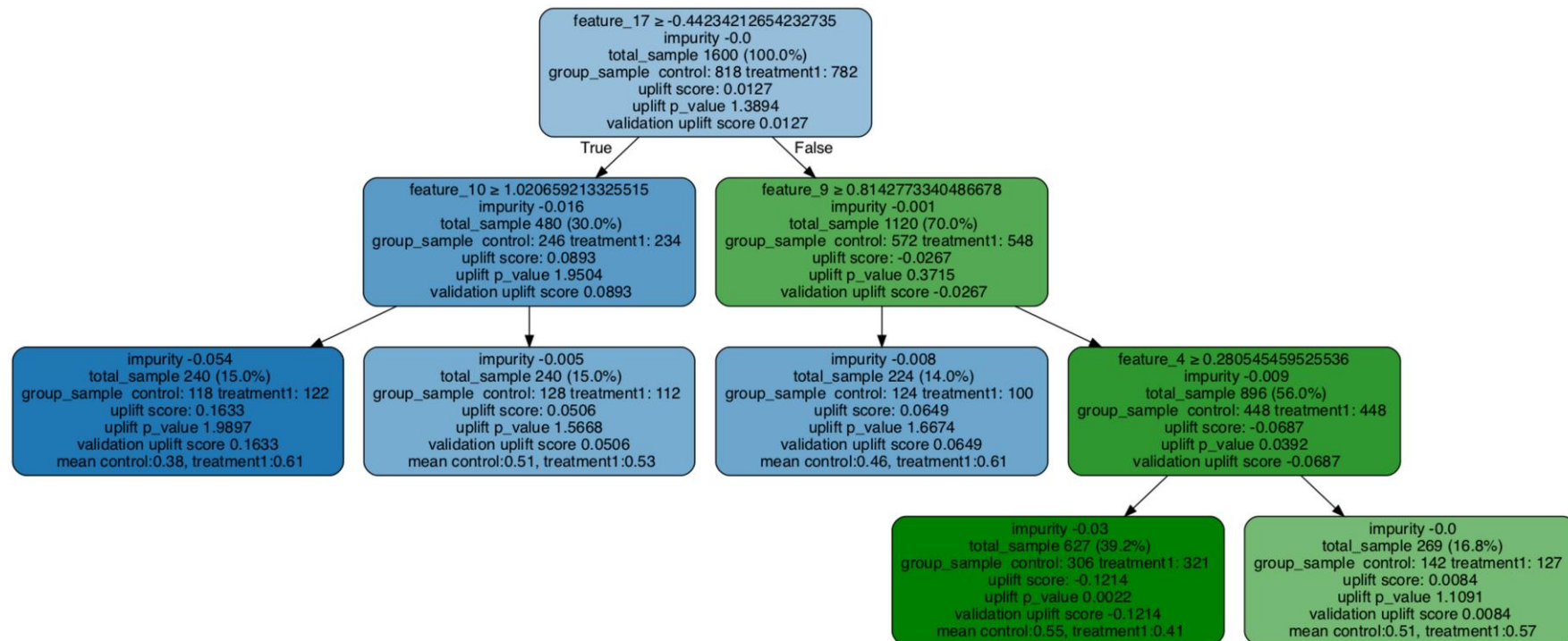


Подходы к решению задачи: Uplift-деревья

Аналогично можно поступить для решения задачи Uplift только выбрав другие критерии:

$$uplift = mean(y_{w=1}) - mean(y_{w=0})$$

$$KL(P:Q) = \sum_{k=left, right} p_k * \log\left(\frac{p_k}{q_k}\right)$$



Метрики качества моделей

Основная сложность — нет правильных ответов. Соответственно у нас нет возможности оценить качество моделей привычным образом.

Uplift curve и Qini curve

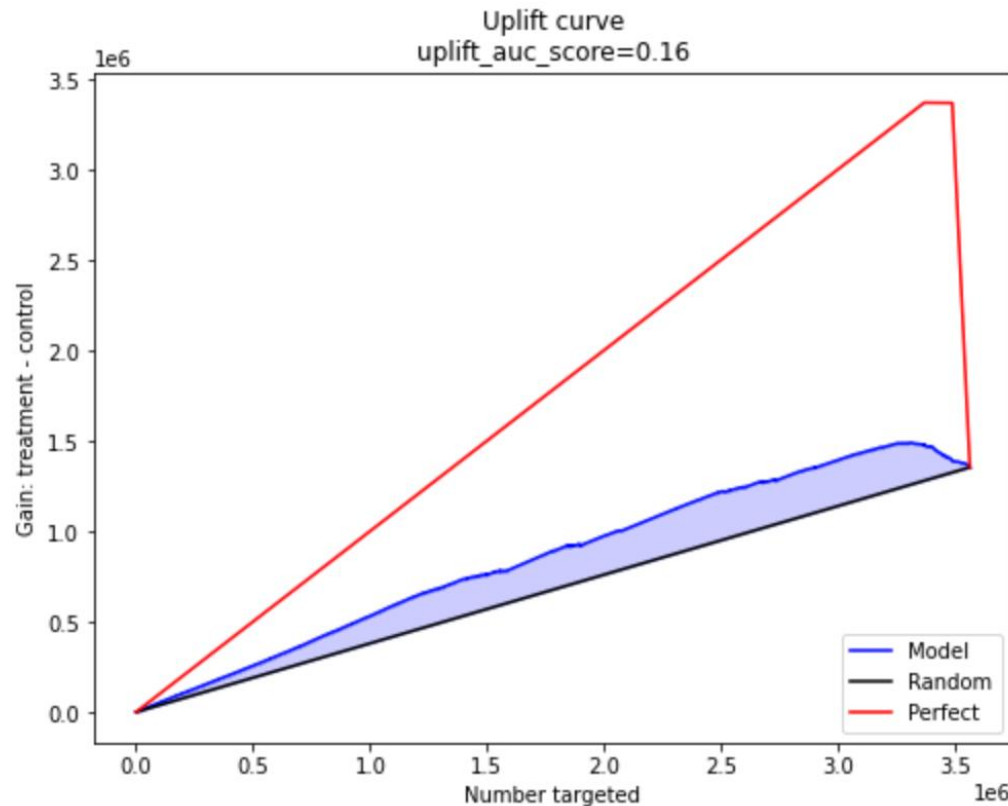
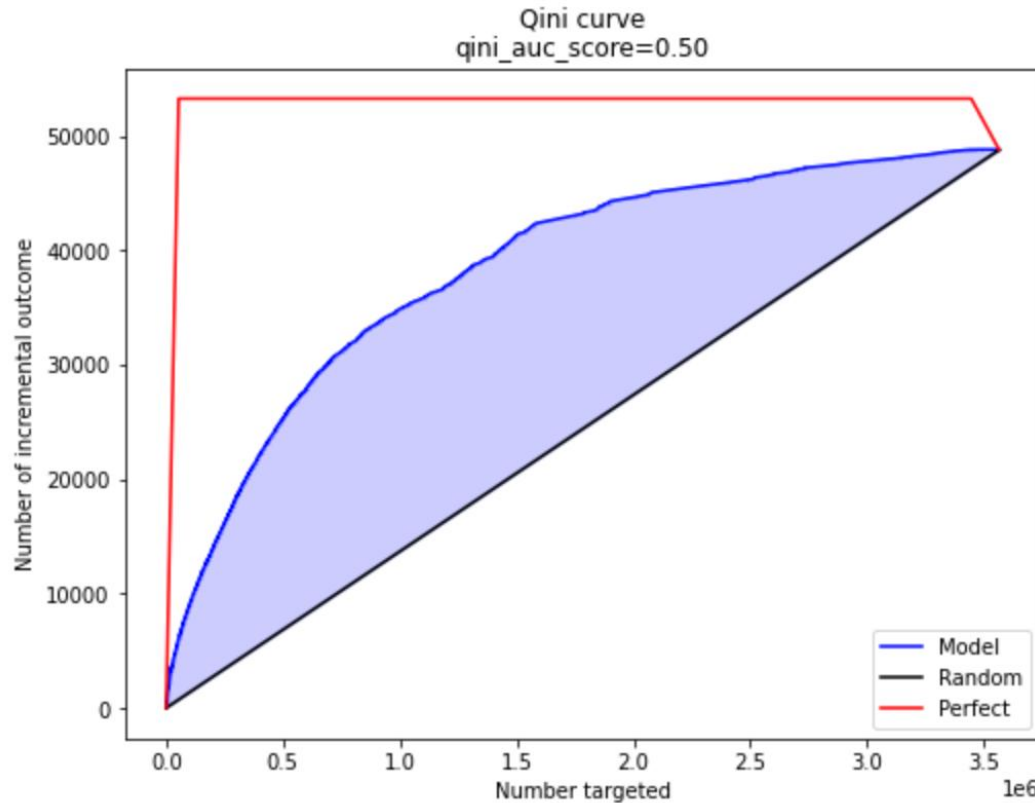
Отсортируем значения выборки по предсказанному uplift. Затем будем идти данной таблице расширяющимся окном, верхняя часть которого фиксирована и находится наверху таблицы, а нижняя часть постепенно увеличивается, соответственно, растёт и количество объектов, попавших в окно. Для каждого размера окна мы вычисляем следующее значение:

$$uplift_curve(t) = ((Y_t^T)/(N_t^T) - (Y_t^C)/(N_t^C)) * (N_t^T + N_t^C)$$

$$qini_curve(t) = Y_t^T - (Y_t^C * N_t^T) / (N_t^C)$$

где Y_t^T - среднее значение таргета (доля покупок) в целевой группе для окна размером t , Y_t^C - среднее значение таргета в контрольной группе для окна размером t , N_t^T - количество клиентов целевой группы, попавшее в окно размером t , N_t^C - количество клиентов контрольной группы, попавшее в окно размером t .

Метрики качества моделей



Фактически, эти кривые показывают то, насколько сильно отличаются друг от друга целевая и контрольная группы (по суммарному, "усредненному" значению таргета). То есть если никаких отличий нет, то и $\text{uplift_curve}(t)$ и $\text{qini_curve}(t)$ будут примерно равны нулю, что соответствует случайной модели.

В качестве числовой оценки качества модели используется площадь под этими кривыми: Qini_auc_score и Uplift_auc_score

Практика – особенности построения Uplift моделей

Спасибо за внимание

tg: @art290790