

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ
ФЕДЕРАЦИИ



ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Выпускная квалификационная работа магистра

Тема: «Разработка алгоритма UpLift моделирования для рекламной кампании»

Выполнил:

студент группы М8О-201М-21
Фейзуллин Кирилл Маратович

Научный руководитель:

к.ф.-м.н., доцент, доценты кафедры 804 МАИ
Платонов Евгений Николаевич

Рецензент:

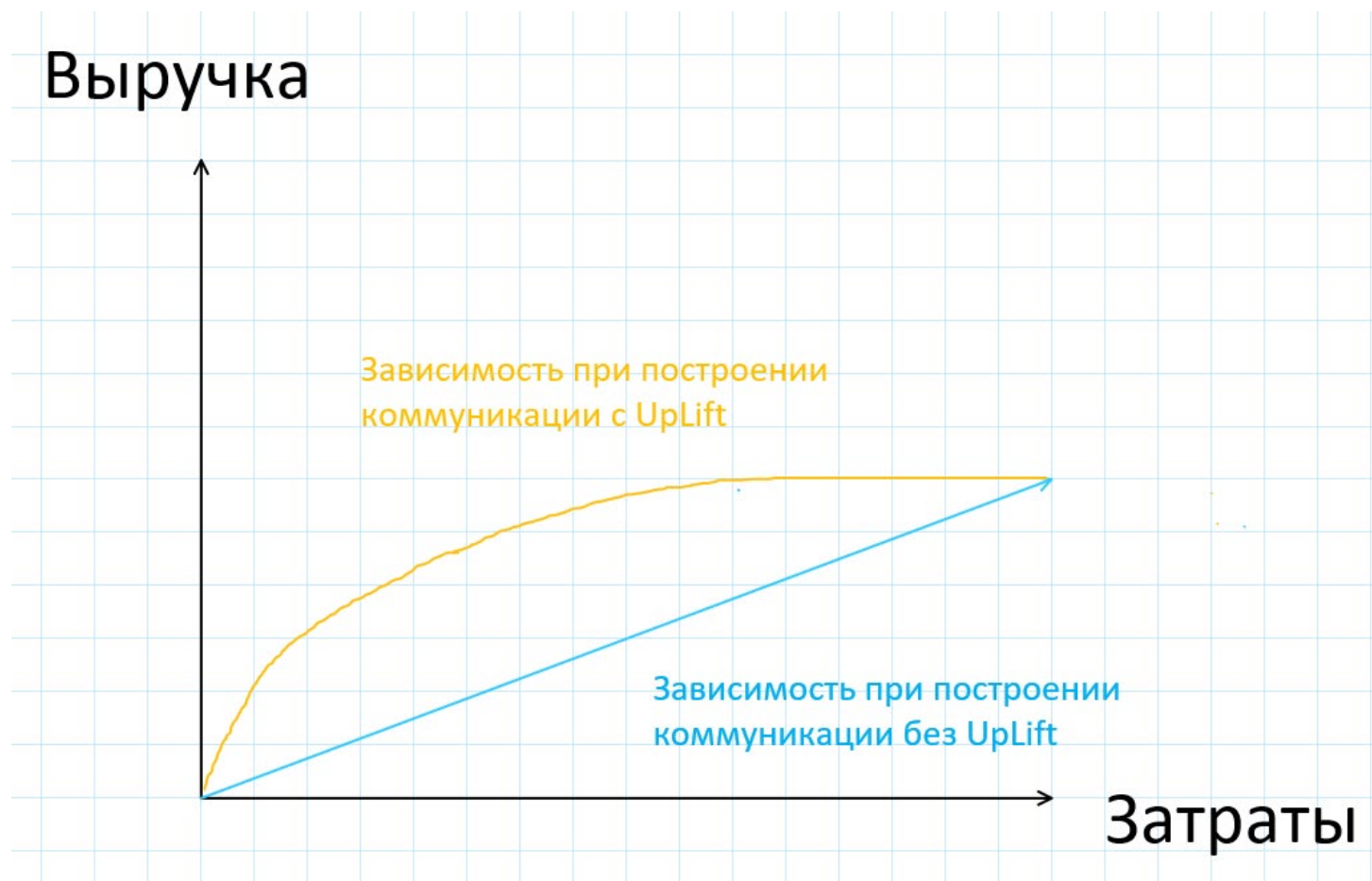
д. ф.-м. н., профессор РАН, ведущий научный сотрудник ФГБУН Института Радиотехники и Электроники
имени В.А. Котельникова РАН
Кузьмин Лев Викторович

Москва, 2023

Актуальность

Данная задача решает проблему оптимизации затрат при большом объеме клиентской базы для коммуникации – пусть коммуникация через СМС на 1 клиента стоит 3 рубля, тогда разовая коммуникация на 1 млн. клиентов стоит уже 3 млн. рублей и появляются вопросы:

- Будет ли от этого экономическая выгода?
- Можно ли получить ту же прибыль от рассылки с меньшими затратами?



Цель работы

- Разработать алгоритм UpLift моделирования для рекламной кампании

Задачи:

- Исследование решений задачи на открытых данных X5-Retail
 - Выбор метрик качества;
 - Выбор используемых моделей и исследование их качества работы;
 - Сравнение полученных результатов.
- Исследование решений задачи на собственных данных ретейл компании косметики и парфюмерии
 - Выбор метрик качества;
 - Выбор используемых моделей и исследование их качества работы;
 - Сравнение полученных результатов.

Описание набора данных X5

Источник данных – открытое соревнование сообщества ODS в партнерстве с “X5 Retail” по UpLift моделированию.

Данные:

- Срез покупок за 4 месяца с детализацией до позиций в чеке.
- Клиентская база объемом около 400 тыс. человек
- Справочник номенклатур позиций в чеке.
- Набор целевых переменных: переменная – флаг воздействия на клиента [0, 1] и переменная – флаг выполнения целевого действия.

	client_id	first_issue_date	first_redeem_date	age	gender
1	000012768d	2017-08-05 15:40:48.00000000	2018-01-04 19:30:07.00000000	45	U
2	000036f903	2017-04-10 13:54:23.00000000	2017-04-23 12:37:56.00000000	72	F
3	000048b7a6	2018-12-15 13:33:11.00000000	1900-01-01 00:00:00.00000000	68	F
4	000073194a	2017-05-23 12:56:14.00000000	2017-11-24 11:18:01.00000000	60	F
5	00007c7133	2017-05-22 16:17:08.00000000	2018-12-31 17:17:33.00000000	67	U

client_id	treatment_flg	target
000012768d	0	1
000036f903	1	1
00010925a5	1	1
0001f552b0	1	1
00020e7b18	1	1

	client_id	transaction_id	TRANSDATE	regular_points_received	express_points_received	regular_points_spent	express_points_spent	AMOUNT	store_id	product_id	QUANTITY	trn_sum_from_iss	trn_sum_from_red
1	2f4980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	4009f09b04	1	5	0
2	2f4980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	f848f9b373	1	38	0
3	2f4980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	34dd2b6d85	1	100	0
4	2f4980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	769a8a92bd	1	51	0
5	2f4980b9bc	1c5d5f6b57	2019-03-09	15,1	0	0	0	2029,89	e87d6aefdc	55e6ba317a	1	184	0

Описание набора данных косметической ретейл компании

Источник данных – исторические данные за 4 месяца до момента коммуникации в косметической ретейл компании
Данные:

- Срез покупок за 4 месяца с детализацией до позиций в чеке.
- Клиентская база объемом около 900 тыс. человек
- Набор целевых переменных: переменная – флаг воздействия на клиента [0, 1] и переменная – флаг выполнения целевого действия.

Results		Messages				
	Дата рассылки	Карта лояльности	treat - параметр наличия СМС	target - целевая переменная	Тип клиента	Канал регистрации
1	2022-11-01	0x6EBD054ACB97355887148DFD14045945	1	1	Новичок	Онлайн
2	2022-11-01	0x09F9A5D3AD73063B770BD0A8A7BB3E7B	1	0	Новичок	Розница
3	2022-11-01	0x539A929BE456EE84074E707E3000CEDB	0	0	Новичок	Розница
4	2022-11-01	0x6432C4BE93BEC38716DC7D7F33C45F2C	1	0	Новичок	Онлайн
5	2022-11-01	0x7E7120709A5DEA46BE0CA5BED4F43735	1	1	Новичок	Онлайн
6	2022-11-01	0x0F38A8435C8557D6A0B259283F28BF7A	1	0	Новичок	Онлайн
7	2022-11-01	0x64C1518274C575F0FA21DCEAF0FBCD64	1	0	Новичок	Онлайн
8	2022-11-01	0x7D7AECC13B11E34CB1923645F3E7722A	1	0	Новичок	Онлайн
9	2022-11-01	0x6C4A553CA03E4C4AD382DD07BAE0F241	1	0	Новичок	Онлайн
10	2022-11-01	0xCC23AFA0E5B2086478A072D51263781D	0	0	Новичок	Онлайн

	Карта лояльности	Дата покупки	Магазин покупки	Касса покупки	Чек покупки	Номенклатура	Сумма	ШТ. товара	Списано бонусов
1	0x4B0EADB857E761E6C4EF48775BC18F94	2022-10-31	AC5	AC6	100076050	CLOR32019	159	1	140
2	0x77844880A0EBDBD5C83280F4BAC27B3B	2022-10-31	AC5	AC6	100076054	LNV013A03	2474	1	300
3	0x97C09F0AE5B5274C590D6AE2BE81C19B	2022-10-31	AC5	AC6	100076060	YSL090008	2122	1	164
4	0xE96996843A03020D3CC8D5A26A74BE8B	2022-10-31	AC5	AC6	100076062	LOTLMP002	602	1	34
5	0xE96996843A03020D3CC8D5A26A74BE8B	2022-10-31	AC5	AC6	100076062	SOD121304	473	1	26
6	0x3A8E58491A38FD31E3ABDEE59E60892E	2022-10-31	AC5	AC6	100076052	CLOR50067	169	1	0
7	0x4528A3C31F85ACF167A1AAA6CA6F01D6	2022-10-31	AC5	AC6	100076053	POI759358	101	1	0
8	0x4528A3C31F85ACF167A1AAA6CA6F01D6	2022-10-31	AC5	AC6	100076053	CLOR20097	349	1	0
9	0x44D8D5E9EBFD96D021D4A02D83B5C897	2022-10-31	AC5	AC6	100076058	CLOR31041	249	1	0
10	0x2F5AC08C159462C583729770CF0E93C7	2022-10-31	AC5	AC6	100076059	ELOR56120	249	1	0

Показатели качества моделирования - 1

- UpLift на первых k – процентах выборки:

$$UpLift_{K\%} = CR_{K\%}(X_{target}) - CR_{K\%}(X_{control}),$$

$$\text{где } CR_{K\%} = \frac{\text{Отклик}_{K\%}}{\text{Размер выборки}_{K\%}}.$$

- Средний взвешенный UpLift (Weighted Average UpLift):

$$WAU = \frac{\sum_{i=1}^k N_i * UpLift_i}{\sum_{i=1}^k N_i},$$

где N_i – размер рабочей выборки на i – м интервале,

$UpLift_i$ – разность конверсий на i – м интервале процентилей (0% – 10%, 11% – 20% и т. д.).

Показатели качества моделирования - 2

- UpLift кривая (UpLift Curve):

$$UC(t) = \left(\frac{N_{target,Y=1}(t)}{N_{target,Y=0,1}(t)} - \frac{N_{control,Y=1}(t)}{N_{control,Y=0,1}(t)} \right) * \left(N_{target,Y=0,1}(t) + N_{control,Y=0,1}(t) \right), \text{ где}$$

$N_{target,Y=0,1}(t)$ – размер всей рабочей группы при всей выборке выборки размера t ,

$N_{target,Y=1}(t)$ –

размер рабочей группы, совершившей целевое действие, при всей выборке размера t .

- Qini кривая:

$$Qini(t) = UC(t) * \frac{N_{target,Y=0,1}(t)}{(N_{target,Y=0,1}(t) + N_{control,Y=0,1}(t))} = \left(\frac{N_{target,Y=1}(t)}{N_{target,Y=0,1}(t)} - \frac{N_{control,Y=1}(t)}{N_{control,Y=0,1}(t)} \right) *$$

$$* N_{target,Y=0,1}(t) = N_{target,Y=1}(t) - N_{control,Y=1}(t) * \frac{N_{target,Y=0,1}(t)}{N_{control,Y=0,1}(t)}$$

Структуры моделей UpLift

Обучающие данные

$$model_1.fit\left(\begin{matrix} Y_1 \\ \vdots \\ Y_m \end{matrix} \middle| \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix} \begin{matrix} W_1 \\ \vdots \\ W_m \end{matrix}\right)$$



Тестовые данные

$$model_1.predict\left(\begin{matrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{matrix} \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}\right) \quad model_1.predict\left(\begin{matrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{matrix} \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right)$$

Обучающие данные

$$model_1.fit\left(\begin{matrix} Y_1 \\ \vdots \\ Y_m \end{matrix} \middle| \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix}, W_i = 1\right) model_2.fit\left(\begin{matrix} Y_1 \\ \vdots \\ Y_m \end{matrix} \middle| \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix}, W_i = 0\right)$$



Тестовые данные

$$\mathbf{UpLift} = model_1.predict\left(\begin{matrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{matrix}\right) \quad model_2.predict\left(\begin{matrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{matrix}\right)$$

Обучающие данные

$$model_1.fit\left(\begin{matrix} Z_1 = Y_1 * W_1 + (1 - Y_1)(1 - W_1) \\ \vdots \\ Z_m = Y_m * W_m + (1 - Y_m)(1 - W_m) \end{matrix} \middle| \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix}\right), Z \in [0; 1]$$



Тестовые данные

$$\mathbf{UpLift} = model_1.predict\left(\begin{matrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{matrix}\right)$$

Обучающие данные

$$model_1.fit\left(\begin{matrix} Z_1 = Y_1 * \frac{W_1 - p}{p * (1 - p)} \\ \vdots \\ Z_m = Y_m * \frac{W_m - p}{p * (1 - p)} \end{matrix} \middle| \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix}\right), Z \in [-2; 0; 2]$$
























Тестовые данные

$$\mathbf{UpLift} = model_1.predict\left(\begin{matrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{matrix}\right)$$

Выбор используемых моделей и исследование их качества работы – X5-Retail.

Номер	Структура	WAU	UpLift на k%	Qini curve AUC	UpLift curve AUC
1	Базовое решение	0,033	0,034	0,000	0,000
2	Решение с одной моделью	0,033	0,032	0,000	0,000
3	Решение с двумя независимыми моделями	0,033	0,053	0,010	0,012
4	Трансформация класса в задачу регрессии	0,033	0,044	0,006	0,006
5	Трансформация класса в задачу регрессии с поиском лучшей модели	0,035	0,070	0,024	0,034

Выбор используемых моделей и исследование их качества работы – собственные данные.

Номер	Структура	UpLift на k%	Qini curve AUC	UpLift curve AUC
1	Базовое решение	 0,0073	 0,0016	 0,0004
2	Решение с одной моделью	 0,0158	 0,0223	 0,0055
3	Решение с двумя независимыми моделями	 0,0144	 0,0167	 0,0042
4	Трансформация класса в задачу классификации	 0,0124	 0,0081	 0,0022
5	Трансформация класса в задачу регрессии	 0,0138	 0,0155	 0,0038
6	Решение с одной моделью с поиском лучшей модели	 0,0233	 0,0543	 0,0136
7	Трансформация класса в задачу регрессии с поиском лучшей модели	 0,0179	 0,0314	 0,0077

Выводы

Наилучшая модель дает прирост на 0.0233. Тогда вероятность покупки с применением UpLift модели составила бы 0.0951, далее найдем экономический прирост: $0.0233 * 473861 * 2500 = 27\,602\,403$ руб.

Таким образом, при сохранении объема расходов на отправку СМС, применение UpLift моделирования в нашем случае принесет 27.6 млн руб. дополнительной выручки при выборке в 473 861 реципиентов.

Помимо этого, наглядно видно, что наилучший алгоритм может отличаться в зависимости от скрытой природ данных (для данных X5 наилучшей оказалось модель регрессии, а для собственных данных простейшая структура с одной независимой моделью)