РЕФЕРАТ

Выпускная квалификационная работа магистра содержит 36 страниц, 30 рисунков, 4 таблиц. Список использованных источников содержит 6 позиций.

UPLIFT, МАШИННОЕ ОБУЧЕНИЕ, ИНКРЕМЕНТАЛЬНЫЙ ОТКЛИК, ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ, РАНЖИРОВАНИЕ КЛИЕНТОВ, ГРАДИЕНТНЫЙ БУСТИНГ, СЛУЧАЙНЫЙ ЛЕС.

Выпускная квалификационная работа магистра посвящена исследованию возможных подходов к решению задачи прогноза инкрементального отклика клиента при получении СМС, при планировании рекламной кампании, с помощью UpLift моделирования.

С ростом глобализации и цифровизации появилась возможность работать с потребительскими данными, активно взаимодействовать с потребителями путем разных акций, особых предложений.

Но стоит взять во внимание, что каждая коммуникация стоит денег. Если клиентская база составляет 1 тыс. клиентов, то при стоимости одного СМС в 1 рубль, коммуникация будет не такой дорогой. Но если увеличить масштаб базы до миллиона или нескольких миллионов, то слепая рассылка всем клиентам подряд станет очень дорогой. Даже если у компании большой оборот выручки, каждая такая коммуникация будет ощутимо сказываться на общем бюджете.

Поэтому коммуникацию можно использовать гораздо более оптимальным способом. Например, совершать коммуникацию с потенциально ушедшим пользователем.

С ростом клиентской базы даже выборочная коммуникация с клиентами будет затратной и следующей задачей является прогнозирование, повлияет ли коммуникация на пользователя.

Содержание

ВВЕДЕНИЕ

В данной выпускной квалификационной работе рассматривается проблема ранжирования клиентов для осуществления коммуникации самым убеждаемым клиентам, которые без той самой коммуникации не совершат целевое действие.

В данной работе решается проблема прогноза инкрементального отклика клиента при планировании коммуникаций с помощью UpLift моделирования методами машинного обучения, где на основании полученного значения будет происходить ранжирование клиентов от самых убеждаемых к самым неприкасаемым, для повышения эффективности коммуникации при сохранении объемов затрат на ее проведение.

Результаты данной работы будут использованы в отделе управления взаимоотношений с клиентами в ретейл компании косметики и парфюмерии.

Появление данной задачи обусловлено желанием проводить нативную коммуникацию только с теми людьми, которым это нужно, чтобы не тратить денежный ресурс в пустую на тех, кому коммуникация не нужна или даже вызовет негативные эмоции и заставит уйти к конкуренту.

Объектом исследования являются клиенты ретейл сети косметики и парфюмерии, которых мы хотим ранжировать для выделения наиболее убеждаемых

Предметом исследования выступает сравнение различных алгоритмов ранжирования методами машинного обучения на двух различных источниках данных.

Цель данной работы - разработка алгоритма UpLift моделирования методами машинного обучения для планирования проведения рекламной кампании.

Основными задачами выпускной квалификационной работы являются:

1. Поиск и обработка информации по объектам исследования;
2. Исследование общих подходов при построении модели UpLift;
3. Построение и обучение моделей UpLift на собственных данных ретейл компании косметики и парфюмерии;
4. Оценка качества построенных моделей с помощью предложенных функционалов качества;
5. Анализ полученных результатов.

По итогам выполнения данной выпускной квалификационной работы поставленные задачи были успешно решены. Результат подтверждает релевантность существующих методов UpLift моделирования с помощью машинного обучения для повышения эффективности рекламной кампании.

Данная работа развивает описанные в [1] идеи прогнозированию эффекта от коммуникации для каждого клиента при планировании рекламной кампании. С помощью показателей качества обучения из [2] и [3] удалось определить наилучший алгоритм для Uplift моделирования из описанных в [1], [4].

ОСНОВНАЯ ЧАСТЬ

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Описание объектов исследования

1.1.1 Исходные данные ретейл компании косметики и парфюмерии

Как и оговаривалось ранее, объектом исследования являются клиенты розничной сети косметики и парфюмерии, по которым имеются исторические данные покупок, заработка и списания бонусов в программе лояльности и многое другое. За источник данных были взяты результаты массовой рассылки СМС в ноябре на 473 861 человек. По истечении недели после рассылки появляется возможность определить целевую переменную (target): 0 – нет покупки в течении недели, 1 – есть покупка в течении недели. И так как нам известно заранее, кому была отослана СМС, а кому нет, очень просто определяется параметр коммуникации (treat): 0 – человек не получал СМС, 1 – человек получил СМС. Для клиентов из эксперимента были рассчитаны покупательские показатели за 4 месяца до момента рассылки, которые будут использованы как обучающие признаки.

Опишем набор данных детальнее. Он состоит из:

- Общая информации о клиентах и целевые переменные для обучения (рисунок 1.1):

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

Рисунок 1.1 – Срез общих анкетные данные клиентов

- История покупок клиентов до коммуникаций (рисунок 1.2):

	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	0x3ABE58491A38FD31E3A6DEE59E60892E	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	0x2F5AC0BC159462C583729770CF0E93C7	2022-10-31	AC5	AC6	100076059	ELOR56120	249	1	0

Рисунок 1.2 – Срез детализации покупок клиентов

1.1.2 Агрегирование данных ретейл компании косметики и парфюмерии

Так как данные для UpLift моделирования находятся в базе SQL Server компании, то было решено и взаимодействовать с ними через реляционный язык запросов T-SQL. Для этого был использован менеджер запросов SQL Management Studio (рисунок 1.3).

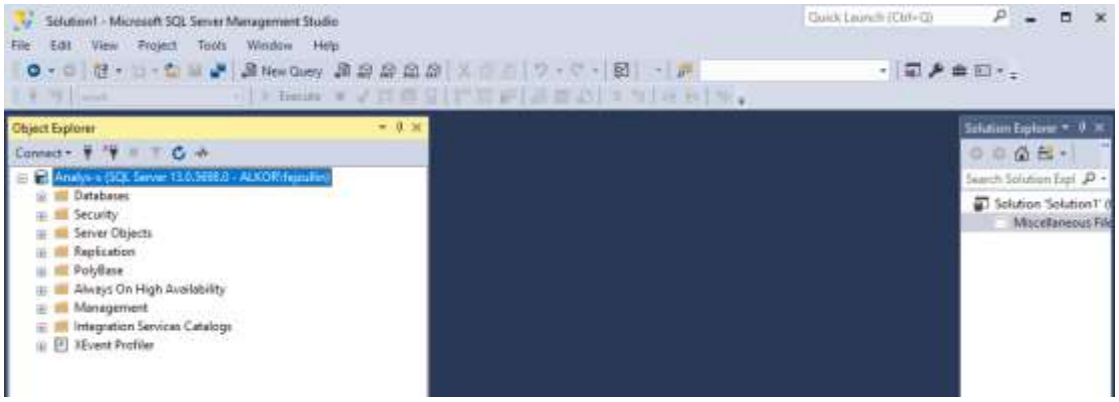

Рисунок 1.3 – Окно среды выполнения SQL запросов

Для моделирования основных обучающих признаков был использован принцип RFM – сегментации [5]. То есть, по покупкам клиентов были определены следующие параметры:

- Частота покупок – количество покупок за расчетный период.
- Период с момента последней покупки.
- Сумма товарооборота с клиента за расчетный период - в

нашем случае возьмем средний чек, так как это стратифицировать клиентов явным образом.

Также была собрана статистика по среднему времени между покупками, минимальном и максимальному интервалу между покупками, а также по трате и заработку бонусов программы лояльности, средняя скидка за счет бонусов, количество покупок и суммы с тратой всех бонусов, количество покупок и суммы с тратой заработанных бонусов, количество покупок и суммы с тратой начисленных в периоды акций бонусов. Вдобавок к этому были учтены и анкетные данные.

Таким образом было получено пространство из 32-ух обучающих признаков (рисунок 1.4):

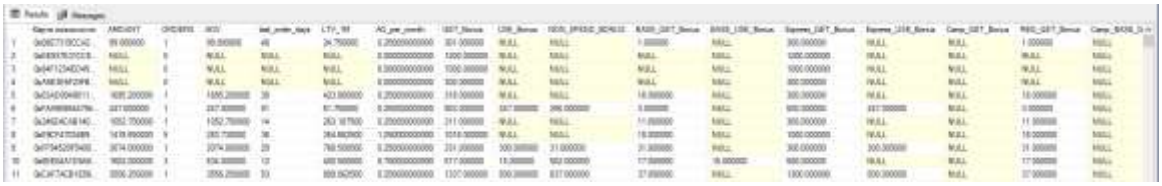

Рисунок 1.4 – Срез агрегированных показателей клиентов

1.1.3 Исходные данные X5-Retail

За источник данных было взято уже завершенное соревнование по UpLift моделированию от российской мега-корпорации X5 Retail Group (ныне X5 Group) на платформе Open Data Science (ODS). Этот набор данных имеет преимущество над ныне существующими в открытом доступе благодаря тому, что это фактически моментальный снимок базы данных компании, во временном интервале за четыре месяца, хранящий в себе транзакции клиентов за соответствующий период, их обезличенные анкетные данные, обезличенный продуктовый справочник с данными по каждому товару сети.

Данное преимущество позволяет самому смоделировать и выделить важные признаки, и получить релевантный опыт работы с живыми, а не синтетическими или уже агрегированными данными.

Опишем набор данных детальнее. Он состоит из:

- Общей информации о клиентах (рисунок 1.5):

	client_id	first_issue_date	first_redeem_date	age	gender
1	000012768d	2017-08-05 15:40:48.0000000	2018-01-04 19:30:07.0000000	45	U
2	000036f903	2017-04-10 13:54:23.0000000	2017-04-23 12:37:56.0000000	72	F
3	000048b7a6	2018-12-15 13:33:11.0000000	1900-01-01 00:00:00.0000000	68	F
4	000073194a	2017-05-23 12:56:14.0000000	2017-11-24 11:18:01.0000000	60	F
5	00007c7133	2017-05-22 16:17:08.0000000	2018-12-31 17:17:33.0000000	67	U

Рисунок 1.5 – Срез анкетных данных клиентов

- Общая информация о товарах на складе (рисунок 1.6):

	product_id	level_1	level_2	level_3	level_4	segment_id	brand_id	vendor_id	netto	is_own_trademark	is_alcohol
1	0003020d3c	c3d3a8e8c6	c2a3ea8d5e	b7cda0ec0c	6376f2a852	123	394a54a7c1	9eaff48661	0,4	0	0
2	0003870676	e344ab2e71	52f13dac0c	d3cfe81323	6dc544533f	105	acd3dd483f	10486c3cf0	0,68	0	0
3	0003ceaf69	c3d3a8e8c6	f2333c90fb	419bc5b424	f6148afbc0	271	f597581079	764e660dda	0,5	0	0
4	000701e093	ec62ce61e3	4202626fcb	88a515c084	48cf3d488f	172	54a90fe769	03c2d70bad	0,112	0	0
5	0007149564	e344ab2e71	52f13dac0c	d3cfe81323	6dc544533f	105	63417fe1f3	f329130198	0,6	0	0

Рисунок 1.6 – Срез справочника товаров

- История покупок клиента до коммуникаций (рисунок 1.7):

	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	294980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	4009f80b04	1	5	0
2	294980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	f848f9b373	1	38	0
3	294980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	34dd2bf6d85	1	100	0
4	294980b9bc	837832dee3	2019-03-06	14,1	0	0	0	1926,68	e87d6aefdc	769a8a92bd	1	51	0
5	294980b9bc	1c5d986b57	2019-03-09	15,1	0	0	0	2029,89	e87d6aefdc	55e6ba317a	1	104	0

Рисунок 1.7 – Срез покупок клиентов

- Целевые переменные для обучения (рисунок 1.8):

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

Рисунок 1.8 – Срез флага коммуникации и целевого действия обучающей выборки

- Данные для теста (рисунок 1.9):

client_id	treatment_flg	target
fffe0abb97	0	0
fffe0ed719	0	1
fffea1204c	0	1
fffeca6d22	1	0
fffff6ce77	0	1

Рисунок 1.9 – Срез флага коммуникации и целевого действия тестовой выборки

1.1.4 Агрегирование данных X5-Retail

Так как данные для UpLift моделирования составляют 4 Гб. в формате csv, что достаточно много для табличных данных самом экономном формате, то было решено взаимодействовать с ними через реляционный язык запросов SQL. Для этого был развернут локальный SQL Server на СУБД MSSQL и с помощью SQL Management Studio были загружены табличные данные.

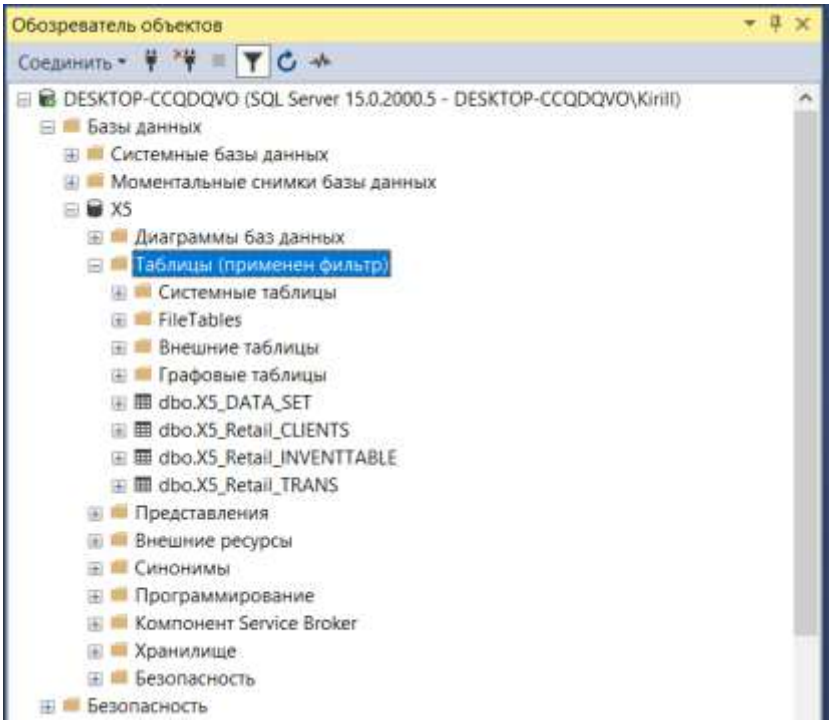

Рисунок 1.10 – Окно среды выполнения SQL запросов со списком используемых таблиц

Через транзакции были выделены наиболее часто покупаемые товары для агрегации их в признаки.

Таким образом были выделены наиболее продаваемые:

- Уровни в иерархии товаров (рисунок 1.11):

	level	level_name	qty	Number
1	1	e344eb2e71	34000656	1
2	1	c3d1e8e8c8	21443244	2
3	1	ec62ce61e3	1648063	3
4	1		8284	4
5	2	ad8b2e17c2	10020239	1
6	2	ed2ad17976	9353348	2
7	2	034ecc0659	5971432	3
8	2	703f1c8ee0	5587036	4
9	2	1d2935be1d	5003892	5
10	3	ca60ed0de2	4889143	1
11	3	b1f8ca38b2	2701359	2
12	3	334b74af37	2349777	3
13	3	a33cc0b2a4	2043846	4
14	3	453bef12c5	1811528	5
15	4	efbcf6d00c	2490149	1
16	4	146717c1b2	2240059	2
17	4	5030a91190	1917686	3
18	4	47fc199714	1454836	4

Рисунок 1.11 – Срез агрегированной иерархии номенклатуры

- Бренды (рисунок 1.12):

	brand	qty	Number
1	4da2dc345f	5815721	1
2	ab230258e9	2886997	2
3		2795792	3
4	037a833d06	2596996	4
5	8281de6bcb	2423379	5

Рисунок 1.12 – Срез агрегированных брендов номенклатуры

- Поставщики (рисунок 1.13):

	vendor	qty	Number
1	43acd80c1a	8622311	1
2	e6af81215a	5688651	2
3	6bc8b3c476	2682034	3
4	63243765ed	1903282	4
5	bf8fc0055c	1584141	5

Рисунок 1.13 – Срез агрегированных поставщиков номенклатуры

- Сегменты товаров (рисунок 1.14):

	segment	qty	Number
1	105	2776209	1
2	230	2701359	2
3	18	2345695	3
4	1	1941509	4
5	9	1824000	5

Рисунок 1.14 – Срез агрегированных сегментов номенклатуры

1.2 Функционалы качества прогноза моделей

1.2.1 UpLift на k – процентах выборки

Так как задача UpLift представляет собой задачу оценки (скор балл) эффекта от коммуникации на реципиента, то нет и истинных ответов. Получается, что не удастся использовать классические метрики, такие как Accuracy и PR AUC, основанные на матрице ошибок, для классификации или среднеквадратичная ошибка для задачи регрессии при трансформации классов.

Самая простая и интуитивно понятная метрика, описанная в [2], особенно для применения в бизнесе и для интерпретации – UpLift на k – процентах выборки.

Допустим, что на коммуникации в компании имеется скромный бюджет, который может обеспечить связь всего с 30% клиентской базы для побуждения к целевому действию. Тогда целью UpLift моделирования будет найти такой алгоритм, который лучше всех максимизирует эффект от коммуникаций на первых 30% клиентов.

Чтобы получить значение этой метрики, нужно ранжировать результат прогноза по убыванию, чтобы отобрать клиентов, на которых коммуникация оказывает наибольший эффект. Далее берется разница между конверсией целевой группы, с которой осуществлялась коммуникация, и конверсией контрольной группы,

которая осталась без коммуникации.

Определяется формулой (1):

$$UpLift_{K\%} = CR_{K\%}\left(X_{target}\right) - CR_{K\%}\left(X_{control}\right), \quad (1)$$

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

Как и сам UpLift, $UpLift_{K\%}$ имеет область значений [-1, 1].

Причем, данную метрику можно рассчитать двумя способами, в зависимости от ранжирования по прогнозу UpLift:

- Сортировка происходит по прогнозу и далее берется разность рабочей и контрольной группы.
- Сортировка происходит внутри каждой группы обособленно и далее берется разность.

Второй вариант имеет более практическое применение, так для оценки эффективности от коммуникаций при рекламных кампаниях, при планировании проведения мероприятий, образуются две однородные выборки – рабочая и тестовая группа.

Для дальнейшего исследования будем оценивать метрику при k = 30%.

1.2.2 Кривая UpLift

Далее определим кривую, которая строится как функция с нарастающим итогом, где для каждой точки задается соответствующий UpLift.

Определяется формулой (2):

$$UC(t) = \left(\frac{N_{target,Y=1}(t)}{N_{target,Y=0,1}(t)} - \frac{N_{control,Y=1}(t)}{N_{control,Y=0,1}(t)}\right) * \left(N_{target,Y=0,1}(t) + N_{control,Y=0,1}(t)\right), \quad (2)$$

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,

аналогично и для контрольной группы - control

Так как данный показатель относительный, он может ввести в заблуждение при интерпретации, а также не будет отражать действительность при неравных пропорция target и control. Поэтому далее опишем более интерпретируемый показатель.

Пример кривой UpLift на рисунке (рисунок 1.15).

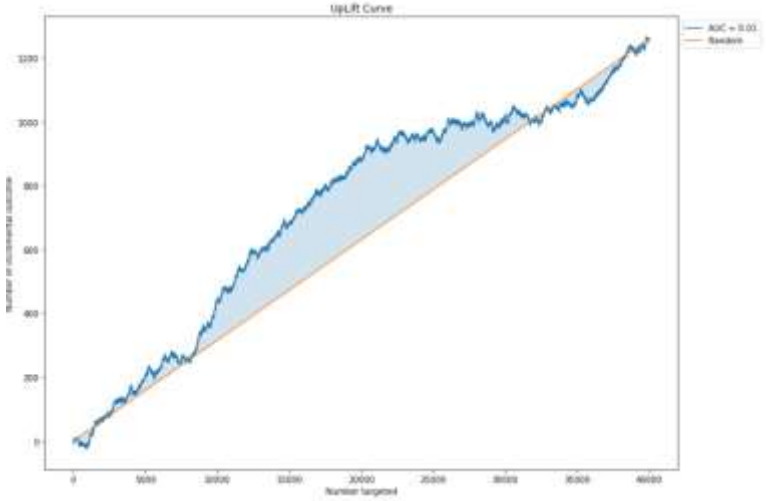

Рисунок 1.15. – Пример кривой UpLift

1.2.3 Кривая QINI

Следующую функцию, описанную в [2], можно выразить через UpLift кривую и получим определение в формуле (3):

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

$$= \left(\frac{N_{target,Y=1}(t)}{N_{target,Y=0,1}(t)} - \frac{N_{control,Y=1}(t)}{N_{control,Y=0,1}(t)}\right) * N_{target,Y=0,1}(t) = \quad (3)$$

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

Данная кривая будет полезна в тех случаях, когда рабочая группа кратно превышает размер контрольной группы, с чем можно столкнуться во время исследования модели при внедрении в бизнес,

когда у компании есть бюджет на произведение коммуникаций со всей клиентской базой, и чтобы не упускать потенциальный доход, контрольная группа выделяется как можно меньше.

Таким образом будет получено инкрементальный эффект от коммуникаций в единицах измерения одного клиента.

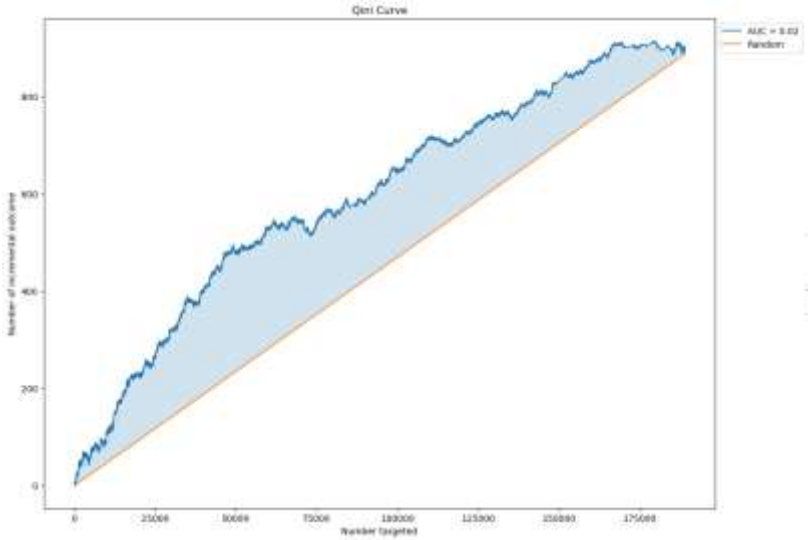

Рисунок 1.16 – Приме кривой QINI

1.3 UpLift моделирование методами машинного обучения

1.3.1 Постановка задачи UpLift

Суть UpLift моделирования в том, чтобы определить, на каких клиентов коммуникация сработает, а на каких нет. Воспользовавшись [1], определим базовые понятия.

Эффект от коммуникации определим как casual effect:

$$\tau_i = Y_i^1 - Y_i^0, \quad (4)$$

где Y_i^1- реакция i – го человека, если коммуникация была, Y_i^0- реакция, если коммуникации не было.

Зная признаковое описание i – го объекта X, можно ввести условный усредненный эффект от воздействия Conditional Average Effect (CATE):

$$CATE(x) = M[Y_i^1 | X_i] - M[Y_i^0 | X_i] \quad (5)$$

Casual effect и CATE можно только оценить, так как одновременно невозможно провести коммуникацию с человеком и не провести. Оценка CATE и является UpLift. Тогда для конкретного объекта он имеет следующее определение:

$$UpLift(x) = M[Y_i | X_i = x, W_i = 1] - M[Y_i | X_i = x, W_i = 0], \quad (6)$$

Где Y_i – наблюдаемая реакция клиента в результате маркетинговой кампании:

$$Y_i = W_I Y_i^1 + (1 - W_i) Y_i^0 = \begin{cases} Y_i^1, \text{если } W_I = 1 \\ Y_i^0, \text{если } W_i = 0 \end{cases} \quad (7)$$

$W_i = 1$, если объект попал в целевую (threatment) группу, в которой была коммуникация,

$W_i = 0$, если объект попал в контрольную (control) группу, в которой коммуникации не было,

$Y_i = 1$, если объект совершил целевое действие,

$Y_i = 0$, если объект не совершил целевое действие

1.3.2 Метод UpLift моделирования с одной независимой моделью

Данный вариант решения из [1] использует переменную W как признак. Тогда обучающий набор данных имеет вид, приведенных в таблице 1.1.

Таблица 1.1 - Пример обучающего набора данных

Обучающие признаки				Целевая
X11	…	X1n	W1	Y1
X21	…	X2n	W2	Y2
……				…
Xm1	…	Xmn	Wm	Ym

С помощью логистической регрессии или подобной модели классификации обучаем модель на данных и после обучения находим

разность вероятностей на тестовой выборке, где в переменной W задаем везде единицы – будто бы была коммуникация, и на той же выборке обрабатываем данные, где в переменной W задаем нули – будто бы единицы не было. Тогда Uplift будет иметь вид:

$$UpLift = P\left(\begin{bmatrix} x_1^1 & \cdots & x_1^n & 1 \\ \vdots & \ddots & \vdots & \\ x_m^1 & \cdots & x_m^n & 1 \end{bmatrix}\right) - P\left(\begin{bmatrix} x_1^1 & \cdots & x_1^n & 0 \\ \vdots & \ddots & \vdots & \\ x_m^1 & \cdots & x_m^n & 0 \end{bmatrix}\right), \quad (8)$$

где P – вероятность целевого действия

1.3.3 Метод UpLift моделирования с двумя независимыми моделями

Второй подход из [1] требует уже обучения двух моделей, одна модель для экспериментальной группы – $P[Y|X = x, W = 1]$, где была коммуникация, вторая модель для контрольной группы $P[Y|X = x, W = 0]$ где коммуникации не было. После обучение моделей на тренировочных выборках, совершается обработка тестовой выборки для каждой модели и за UpLift берется так же разность двух вероятностей:

$$UpLift = P_1\left(\begin{bmatrix} x_1^1 & \cdots & x_1^n & 1 \\ \vdots & \ddots & \vdots & \\ x_m^1 & \cdots & x_m^n & 1 \end{bmatrix}\right) - P_2\left(\begin{bmatrix} x_1^1 & \cdots & x_1^n & 0 \\ \vdots & \ddots & \vdots & \\ x_m^1 & \cdots & x_m^n & 0 \end{bmatrix}\right), \quad (9)$$

где P_1 – вероятность целевого действия первой модели, а P_2 – вероятность целевого действия первой модели

1.3.4 Метод трансформации класса (задача классификации)

В данном методе из [1] мы вернемся снова к единой модели, но теперь преобразуем коммуникационную переменную и целевую переменную в одну следующим образом:

$$Z_i = Y_i * W_i + (1 - Y_i)(1 - W_i), \quad (10)$$

где Y_i-целевая переменная,W_i -коммуникационная переменная.

Тогда трансформированный класс будет иметь следующие значения:

$$Z_i = \begin{cases} 1 \text{ при } W_i = 1;\ Y_i = 1 \\ 0 \text{ при } W_i = 0;\ Y_i = 1 \\ 0 \text{ при } W_i = 1;\ Y_i = 0 \\ 1 \text{ при } W_i = 0;\ Y_i = 0 \end{cases} \quad (11)$$

Тогда UpLift будет определяться следующим образом по формуле (12):

$$UpLift = P\left(\begin{bmatrix} x_1^1 & \cdots & x_1^{\mathrm{n}} \\ \vdots & \ddots & \vdots \\ x_m^1 & \cdots & x_{\mathrm{m}}^{\mathrm{n}} \end{bmatrix}\right), \quad (12)$$

где P – вероятность выполнения закодированного целевого действия

1.3.5 Метод трансформации класса (задача регрессии)

В данном методе мы преобразуем коммуникационную переменную и целевую переменную в одну следующим образом:

$$Z_i = Y_i * \frac{W_i - p}{p*(1-p)}, \quad (13)$$

где Y_i – целевая переменная, $W_i -$ коммуникационная переменная, $p = P(W = 1) = \frac{N_{target}}{N}$ – вероятность принадлежности к целевой группе.

В нашем случае, $p = 0.5$. Тогда трансформированный класс будет иметь следующие значения:

$$Z_i = \begin{cases} 2, \text{при } W_i = 1;\ Y_i = 1 \\ 0, \text{при } W_i = 0{,}1;\ Y_i = 1 \\ -2, \text{при } W_i = 0; Y_i = 1 \end{cases} \quad (14)$$

Тогда UpLift будет определяться следующим образом по формуле (15):

$$UpLift = R\left(\begin{bmatrix} x_1^1 & \cdots & x_1^{\mathrm{n}} \\ \vdots & \ddots & \vdots \\ x_m^1 & \cdots & x_{\mathrm{m}}^{\mathrm{n}} \end{bmatrix}\right), \quad (15)$$

где R – регрессионное значение закодированного целевого действия

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Экспериментальная установка

Исследование методов UpLift моделирования с помощью машинного обучения реализовано на высокоуровневом языке программирования Python, с использованием библиотек scikit-learn, scikit-uplift, CatBoost.

Для сравнения методов моделирования используется модель градиентного бустинга с базовыми параметрами, реализованный в библиотеке CatBoost.

Чтобы избежать ложных выводов по результатам работы модели на тестовом множестве, в исследовании используется кросс валидация [6] с разбиением выборки на 5 долей. По итогу кросс валидации будут браться средние показатели качества обучения, на основе которых и будет сравнение. Иллюстрация работы кросс валидации на рисунке ниже (рисунок 2.1).

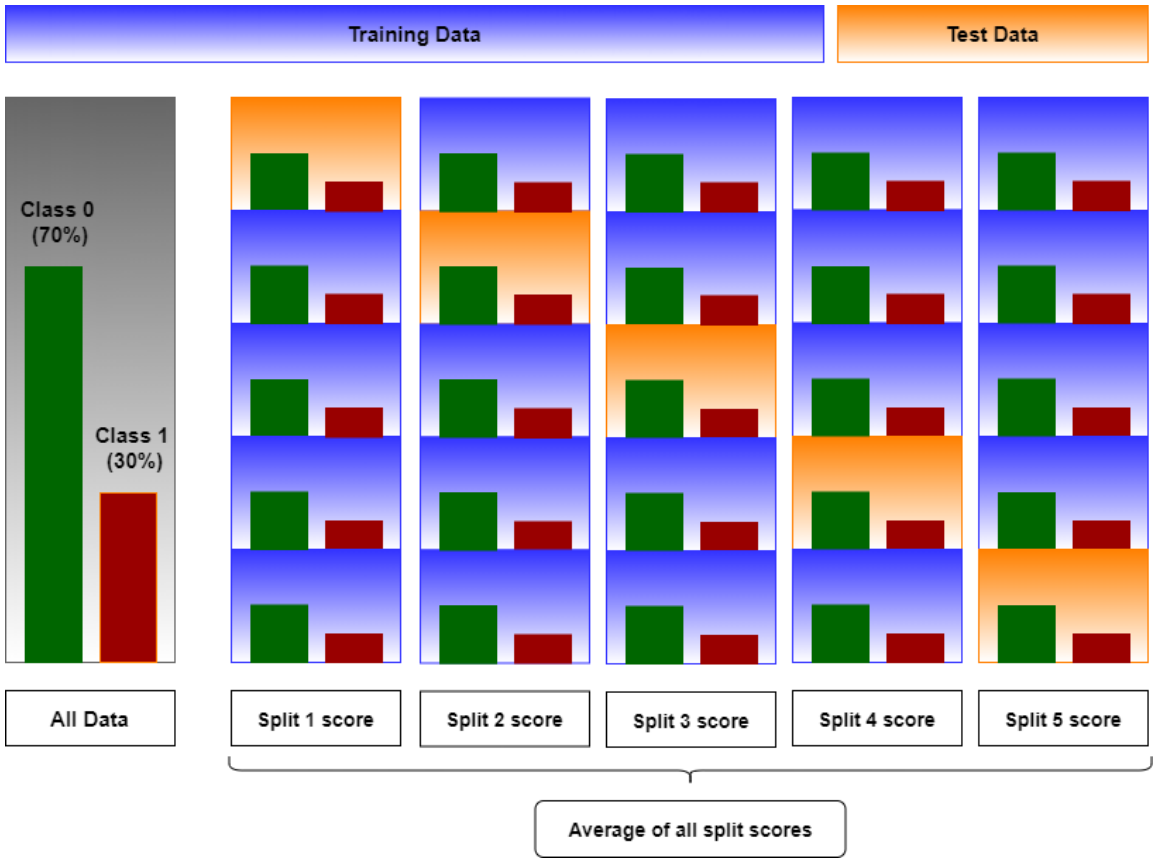

Рисунок 2.1 - Схема кросс валидации

2.2 Базовая модель

Перед проведением экспериментов следует определить базовую модель, от функционала качества которой нужно будет отталкиваться. Так как базовая модель предполагает слепое прогнозирование без обработки пространства признаков, в нашем случае подойдет равномерная случайная величина, распределенная от -1 до 1.

По итогам такого моделирования на собственных получаем следующие значения метрик:

- $\text{UpLift}_{30\%}$= 0.0073
- Qini curve AUC = -0.0016
- UpLift curve AUC = -0.0004

По итогам такого моделирования на данных X5-Retail получаем следующие значения метрик:

- $\text{UpLift}_{30\%}$= 0.0341
- Qini curve AUC = 0
- UpLift curve AUC = 0.

2.3 Моделирование с одной моделью

Самое простое и понятное решение. На тренировочной выборке обучаем любую модель бинарной классификации по всем обучающим признакам, включая коммуникационную переменную.

Далее для тестовой выборки задаем коммуникационную переменную равную 1 и определяем прогноз вероятности, что объект совершит целевое действие.

Далее для тестовой выборки задаем коммуникационную переменную равную 0 и снова определяем прогноз вероятности, что объект совершит целевое действие.

После этого берется разность вероятностей при наличии

коммуникации и при отсутствии, что и будет значением UpLift.

По итогам такого моделирования на собственных данных получаем следующие усредненные метрики:

- $\mathrm{UpLift}_{30\%}$= 0.0158
- Qini curve AUC = 0.0223
- UpLift curve AUC = 0.0055

По итогам такого моделирования на данных X5-Retail получаем следующие значения метрик:

- $\mathrm{UpLift}_{30\%}$= 0.0319
- Qini curve AUC = 0
- UpLift curve AUC = 0.

По итогу кросс валидации на собственных имеются два типа событий:

- Когда моделирование дает наилучший UpLift (рисунок 2.2).

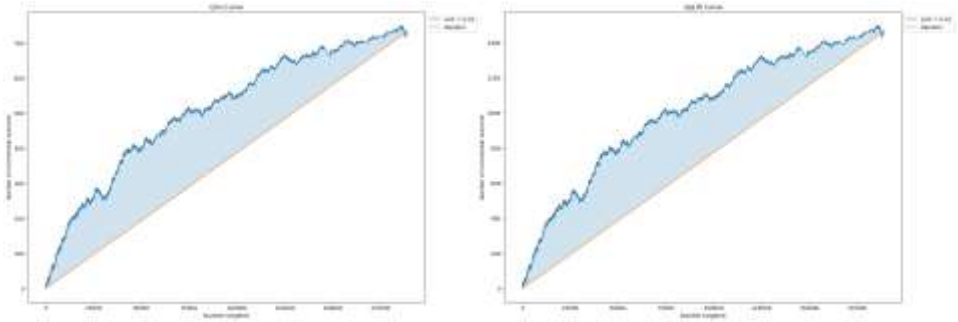

Рисунок 2.2 - Графики кривой QINI и UpLift для результатов моделирования с одной моделью в лучшем случае

- Когда моделирование дает наихудший UpLift (рисунок 2.3).

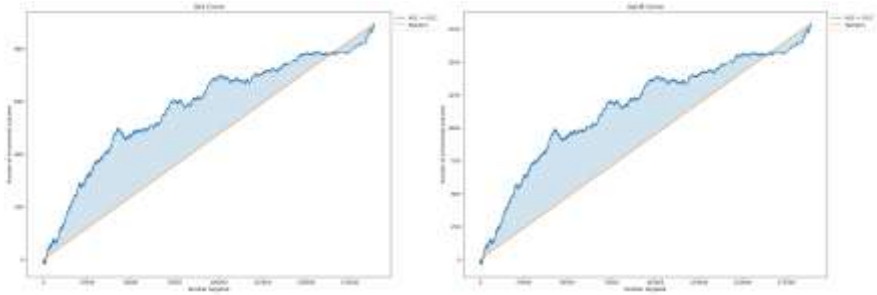

Рисунок 2.3 - Графики кривой QINI и UpLift для результатов моделирования с одной моделью в худшем случае

2.4 Моделирование с двумя независимыми моделями

Метод представляет собой обучение двух независимых моделей на тренировочных данных, где одна модель обучается на целевой группе, а вторая обучается на контрольной. Далее на тестовых данных прогнозируется вероятность выполнения целевого действия для одной и для второй модели и берется их разность.

Но тут сразу возникает нюанс, что при отсутствии равного объема целевой и контрольной группы, модели не будут иметь одинаковую полноту обучения. Но в нашем случае этого происходить не будет, так как рабочая и тестовая группа равного объема. Однако стоит учитывать этот нюанс при заготовке исторических данных для обучения моделей машинного обучения, так как если этого не сделать, то результаты эксперимента могут быть не объективными.

По итогам моделирования на собственных данных получены следующие усредненные метрики:

- $\text{UpLift}_{30\%}$= 0.0144
- Qini curve AUC = 0.0167
- UpLift curve AUC = 0.0042

По итогам моделирования на данных X5-Retail получены следующие усредненные метрики:

- $\text{UpLift}_{30\%}$= 0.0534
- Qini curve AUC = 0.01
- UpLift curve AUC = 0.012

По итогу кросс валидации на собственных данных имеются два типа событий:

- Когда моделирование дает наилучший UpLift (рисунок 2.4):

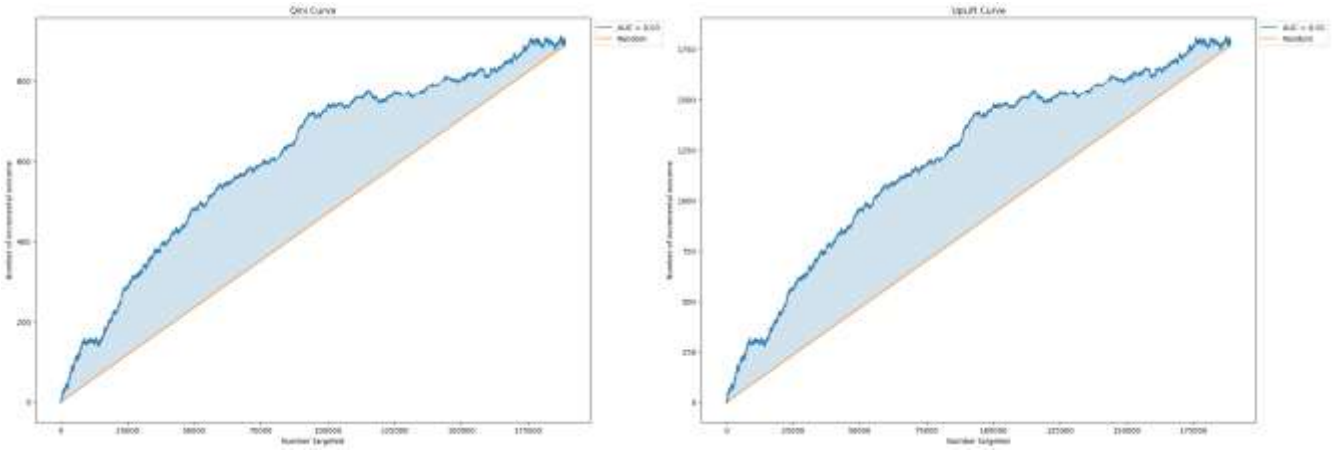

Рисунок 2.4 - Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

- Когда моделирование дает наихудший UpLift (рисунок 2.5):

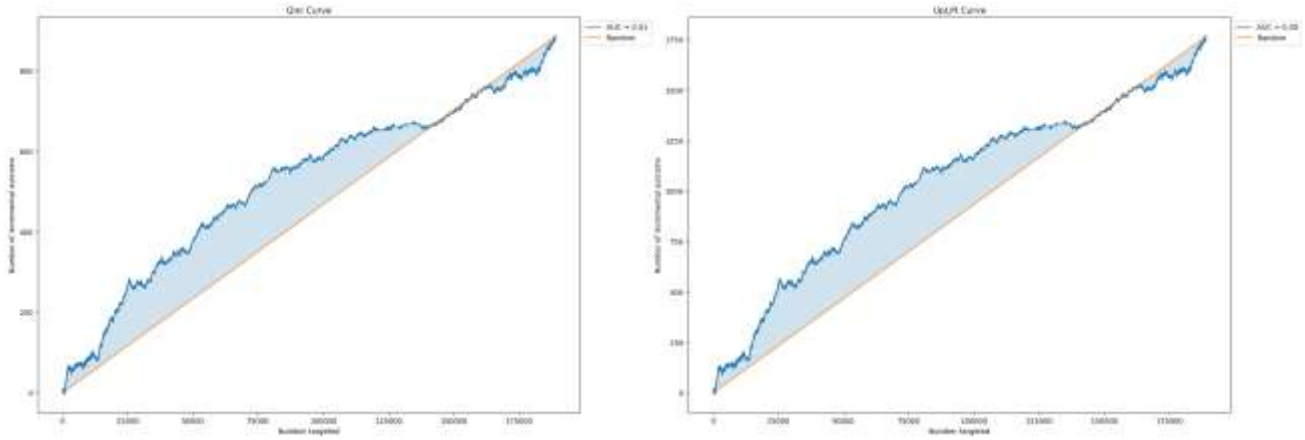

Рисунок 2.5 - Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

Так же стоит добавить, что поведение показателей качества обучения на тестовой выборке в 4 из 5 итераций кросс валидации выглядит как на (рисунок 2.5), что говорит об ухудшении качества обучения – о чем и сигнализируют усредненные показатели $\mathrm{UpLift}_{30\%}$, Qini curve AUC, UpLift curve AUC.

2.5 Метод трансформации класса (задача классификации)

Напомню, как и описывал в теории ранее, в данном методе мы вернемся снова к единой модели, но теперь преобразуем коммуникационную переменную и целевую переменную в одну по формуле (10)

Тогда трансформированный класс будет иметь следующие значения, описанные по формуле (11)

По результатам моделирования на собственных данных были получены следующие усредненные результаты:

- $\text{UpLift}_{30\%}$= 0.0124
- Qini curve AUC = 0.0081
- UpLift curve AUC = 0.0022

По итогу кросс валидации на собственных данных имеются два типа событий:

- Когда моделирование дает наилучший UpLift (рисунок 2.6).

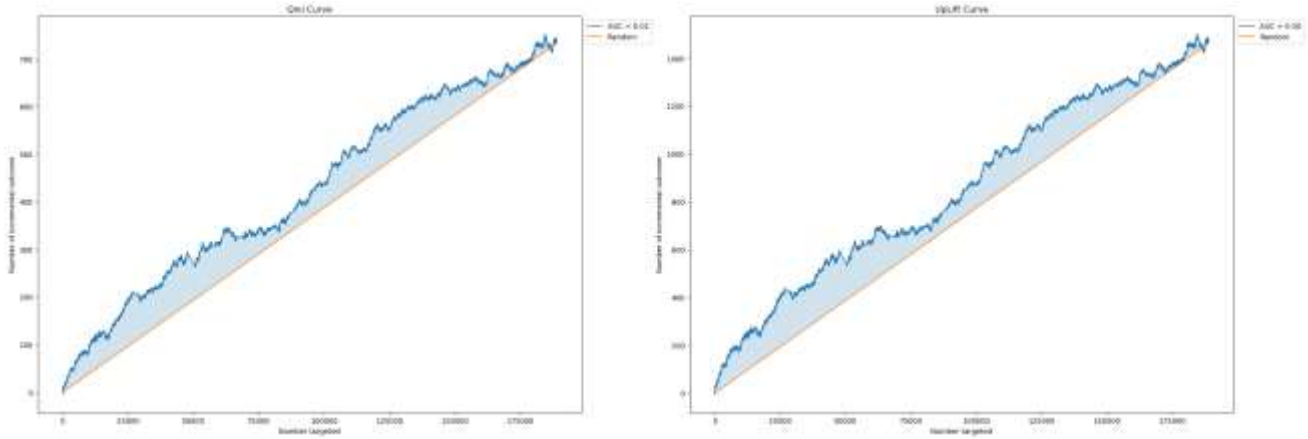

Рисунок 2.6 - Графики кривой QINI и UpLift для результатов моделирования с трансформацией класса с переходом к задаче классификации в лучшем случае

- Когда моделирование дает наихудший UpLift (рисунок 2.7):

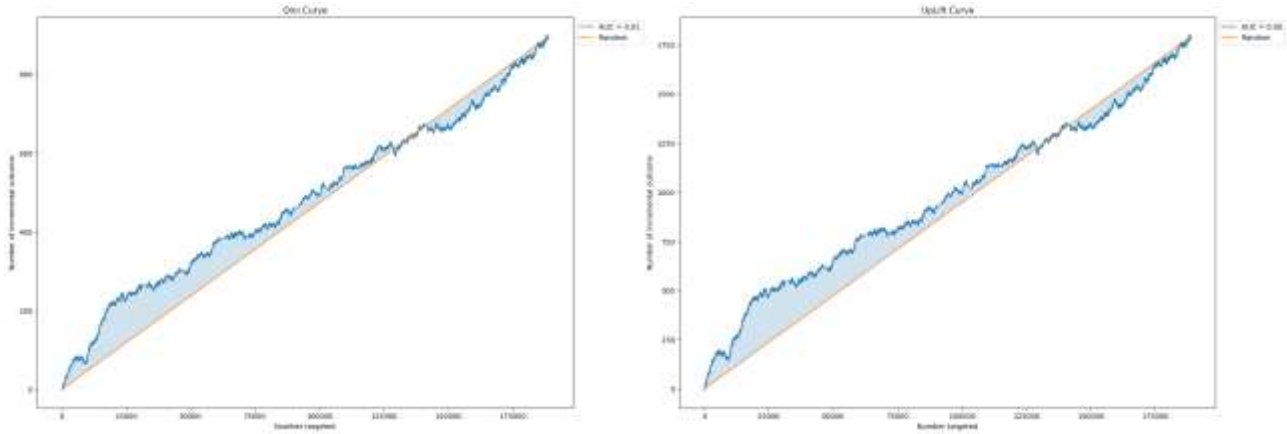

Рисунок 2.7 - Графики кривой QINI и UpLift для результатов моделирования с трансформацией класса с переходом к задаче классификации в худшем случае

Метод трансформации класса в задаче классификации показывает еще более худшие показатели качества обучения, чуть ли не в 2 раза хуже, чем в моделировании с двумя независимыми моделями.

2.6 Метод трансформации класса (задача регрессии)

В данном методе мы вернемся снова к единой модели, но теперь преобразуем коммуникационную переменную и целевую переменную в одну по формуле (13)

В нашем случае, $p = 0.5$. Тогда трансформированный класс будет определен по формуле (14)

Далее произведем переход к задаче регрессии для однозначной интерпретации прогноза.

По результатам моделирования на собственных данных были получены следующие усредненные результаты:

- $\text{UpLift}_{30\%}$= 0.0138
- Qini curve AUC = 0.0155
- UpLift curve AUC = 0.0038

По результатам моделирования на собственных данных были

получены следующие усредненные результаты:

- $UpLift_{30\%}$= 0.0441
- Qini curve AUC = 0.006
- UpLift curve AUC = 0.006

По итогу кросс валидации на собственных данных имеются два типа событий:

- Когда моделирование дает наилучший UpLift (рисунок 2.8).

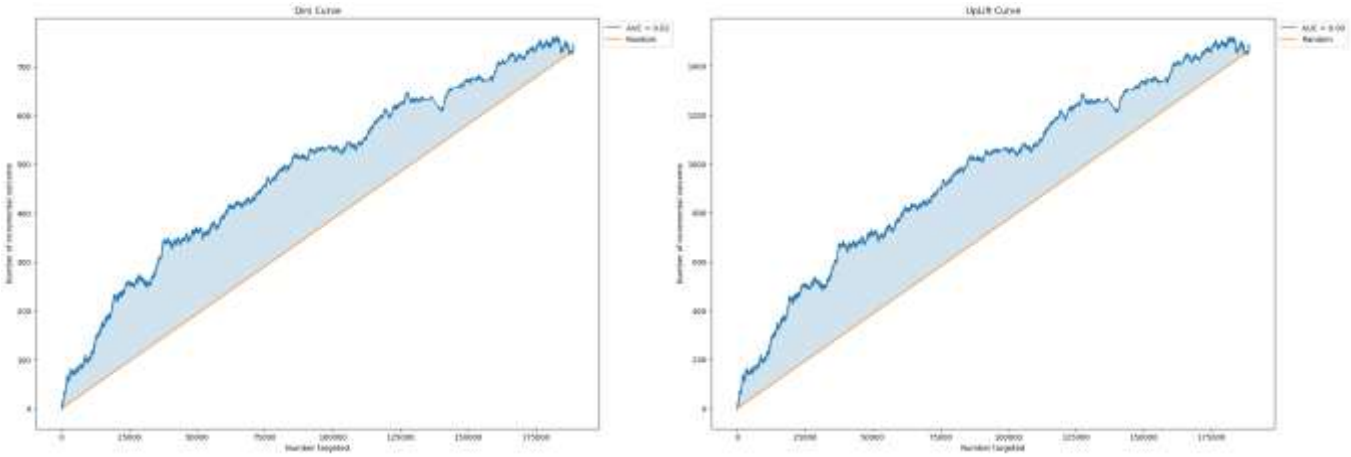

Рисунок 2.8 - Графики кривой QINI и UpLift для результатов моделирования с трансформацией класса с переходом к задаче регрессии в лучшем случае

- Когда моделирование дает наихудший UpLift (рисунок 2.9).

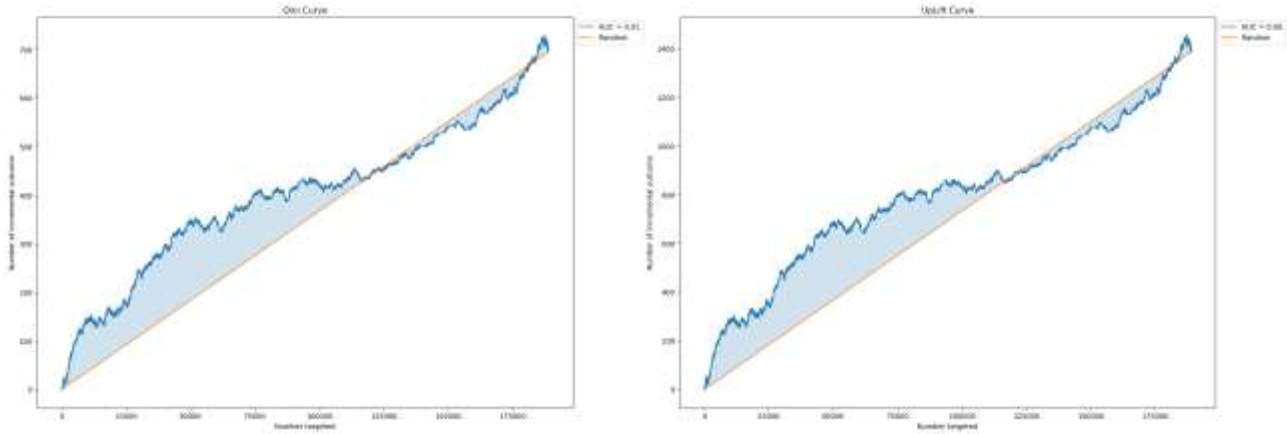

Рисунок 2.9 - Графики кривой QINI и UpLift для результатов моделирования с трансформацией класса с переходом к задаче регрессии в худшем случае

2.7 Исследований архитектур моделей машинного обучения

2.7.1 Поиск лучшей архитектуры для задачи классификации

Так как UpLift моделирование напрямую зависит от качества обучения на наших данных, чтобы максимизировать наши результаты, найдем наилучшую структуру модели классификации клиента, где целевым признаком будет факт покупки (target) и найдем для нее целевые показатели.

Сравнение структур моделей будет происходить с помощью библиотеки evalml, которая содержит внутри себя уже весь реализованный функционал.

По итогам поиска по 13-ти моделей, наилучшие показатели имеет уже использованный ранее градиентный бустинг из библиотеки Яндекс CatBoost. Лучшие результаты в таблице 2.1.

Таблица 2.1 – Результаты автоматического поиска лучшей модели в задаче классификации для собственных данных

Номер	**pipeline_name**	**validation_score**	**percent_better_baseline**
1	Stacked Ensemble Classification Pipeline	0,415	4047%
2	Random Forest Classifier w/ Label Encoder + Replace Nullable Types Transformer + Imputer + Undersampler	0,415	4046%
3	LightGBM Classifier w/ Label Encoder + Replace Nullable Types Transformer + Imputer + Undersampler + Select Columns Transformer	0,406	3965%

Далее взяли лучший PipeLine – ансамбль из моделей: Логистическая Регрессия, Случайный Лес, Дерево Решений, Градиентный бустинг LigthGBM, Расширенные Деревья (Extra Trees), Градиентный бустинг CatBoost, Градиентный бустинг XGBoost. И модель классификации, обрабатывающая результаты ансамбля – ElasticNet.

Далее возьмем эту наилучшую архитектуру и применим ее для моделирования UpLift с одной независимой моделью, описанную в пункте 1.3.2 и найдем усредненные показатели функционалов качества.

По результатам моделирования были получены следующие усредненные результаты:

- $\text{UpLift}_{30\%}$= 0.0233
- Qini curve AUC = 0.0543
- UpLift curve AUC = 0.0136

По итогу кросс валидации имеются два типа событий:

- Когда моделирование дает наилучший UpLift (рисунок 2.10).

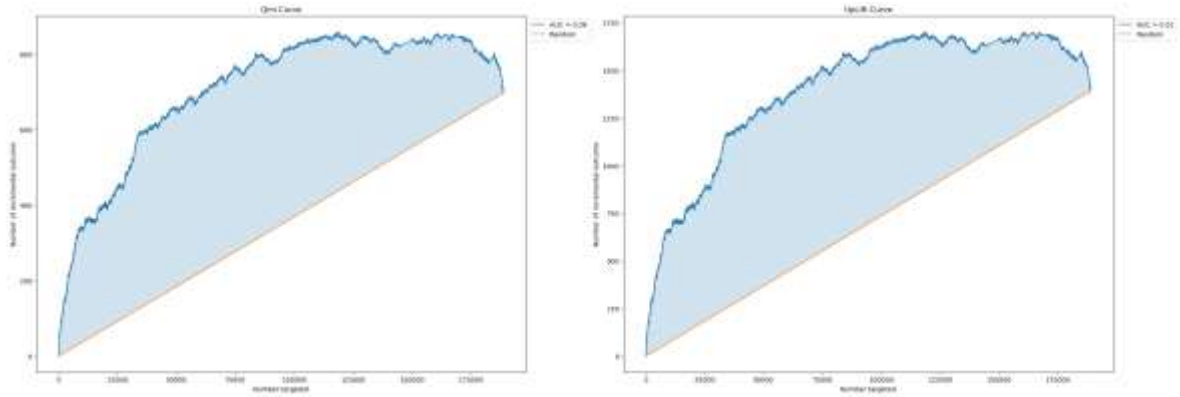

Рисунок 2.10 - Графики кривой QINI и UpLift для результатов моделирования с одной моделью в лучшем случае

- Когда моделирование дает наихудший UpLift (рисунок 2.11).

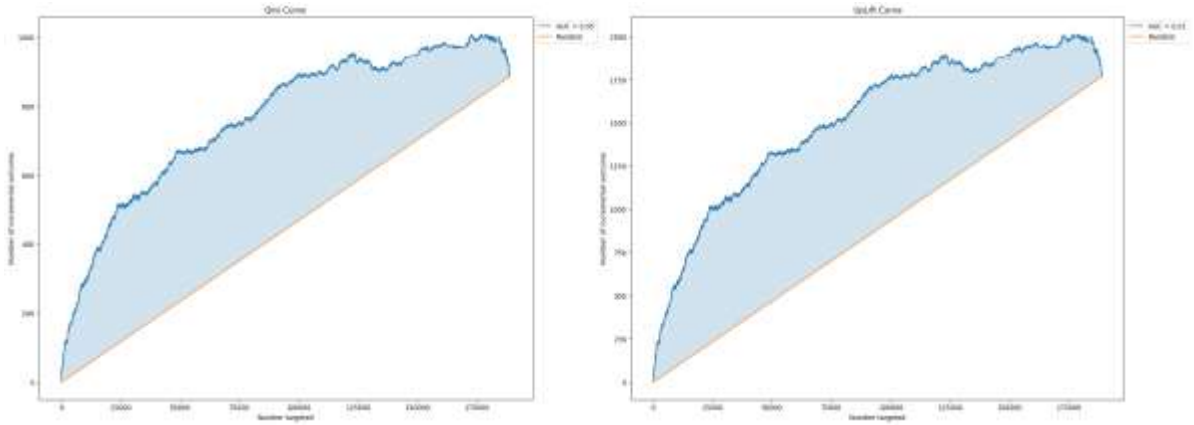

Рисунок 2.11 - Графики кривой QINI и UpLift для результатов моделирования с одной моделью в худшем случае

2.7.2 Поиск лучшей архитектуры для задачи регрессии

Так как по результатам подходов наилучшие имеет метод трансформации классов с переходом к задаче регрессии, то возникает вопрос – какая модель позволяет получить наилучший результат для нашей задачи.

Если считать, что наши целевые переменные достоверные, то косвенно оценивать качество моделей для сравнения можно и с помощью среднеквадратичной ошибки. Ведь та модель, которая лучше всего обучиться на тренировочных данных и тестовых данных и должна потенциально иметь наилучший UpLift на практике.

Сравнение структур моделей будет происходить с помощью библиотеки evalml, которая содержит внутри себя уже весь реализованный функционал.

По итогам поиска по 11-ти моделям, наилучшие показатели имеет уже использованный ранее градиентный бустинг из библиотеки Яндекс CatBoost. Лучшие результаты в таблице 2.2.

Таблица 2.2 – Результаты автоматического поиска лучшей модели в задаче регрессии на собственных данных

Номер	**pipeline_name**	**validation_score**	**percent_better_baseline**
1	CatBoost Regressor w/ Replace Nullable Types Transformer + Imputer + Select Columns Transformer	0,27092	0,3873%
2	Elastic Net Regressor w/ Replace Nullable Types Transformer + Imputer + Standard Scaler + RF Regressor Select From Model	0,27093	0,2225%
3	Mean Baseline Regression Pipeline	0,27093	0,0000%

Далее взяли лучший PipeLine: регрессионная модель градиентного бустинга от Яндекс - CatBoost, с выбором наиболее значимых для модели параметров.

Далее возьмем эту наилучшую архитектуру и применим ее для моделирования UpLift с одной независимой моделью, описанную в пункте 1.3.5 и найдем усредненные показатели функционалов качества.

Для данных X5-Retail так же по итогам поиска по 11-ти моделям, наилучшие показатели имеет уже использованный ранее градиентный бустинг из библиотеки Яндекс CatBoost. Лучшие результаты в таблице 2.3.

Таблица 2.3 – Результаты автоматического поиска лучшей модели в задаче регрессии на данных X5-Retail

Номер	**pipeline_name**	**validation_score**	**percent_better_baseline**
1	CatBoost Regressor w/ Replace Nullable Types Transformer + Imputer + Select Columns Transformer	1,574504	0,001993
2	Mean Baseline Regression Pipeline	1,574535	0
3	Elastic Net Regressor w/ Replace Nullable Types Transformer + Imputer + Standard Scaler + Select Columns Transformer	1,574535	0

По результатам моделирования на собственных данных были получены следующие усредненные результаты:

- $\mathrm{UpLift}_{30\%}$= 0.0179
- Qini curve AUC = 0.0314
- UpLift curve AUC = 0.0077

По результатам моделирования на данных X5-Retail были получены следующие усредненные результаты:

- $\mathrm{UpLift}_{30\%}$= 0.0699
- Qini curve AUC = 0.024
- UpLift curve AUC = 0.034

По итогу кросс валидации на собственных данных имеются два типа событий:

- Когда моделирование дает наилучший UpLift (рисунок 2.12).

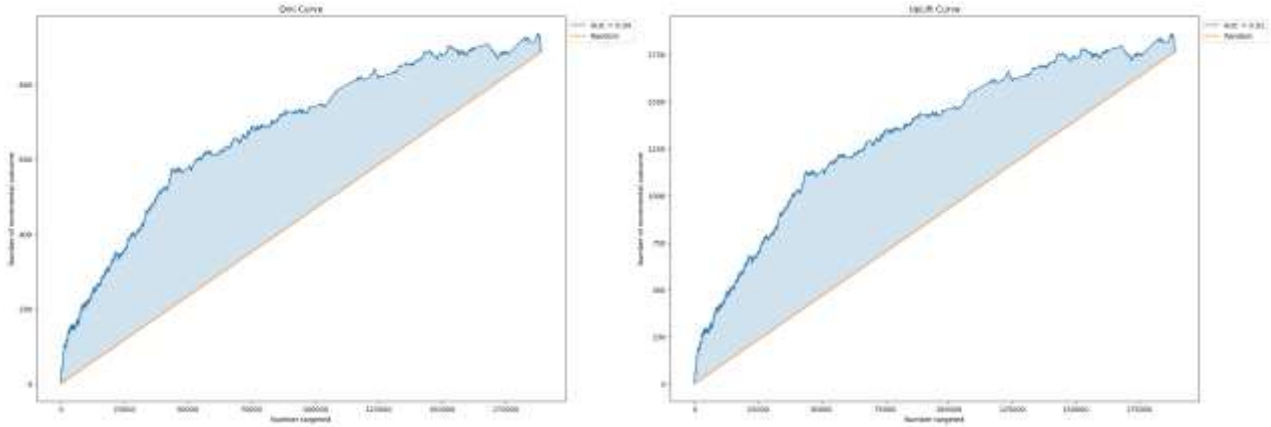

Рисунок 2.12 - Графики кривой QINI и UpLift для результатов моделирования с трансформацией класса с переходом к задаче регрессии в лучшем случае

- Когда моделирование дает наихудший UpLift (рисунок 2.13).

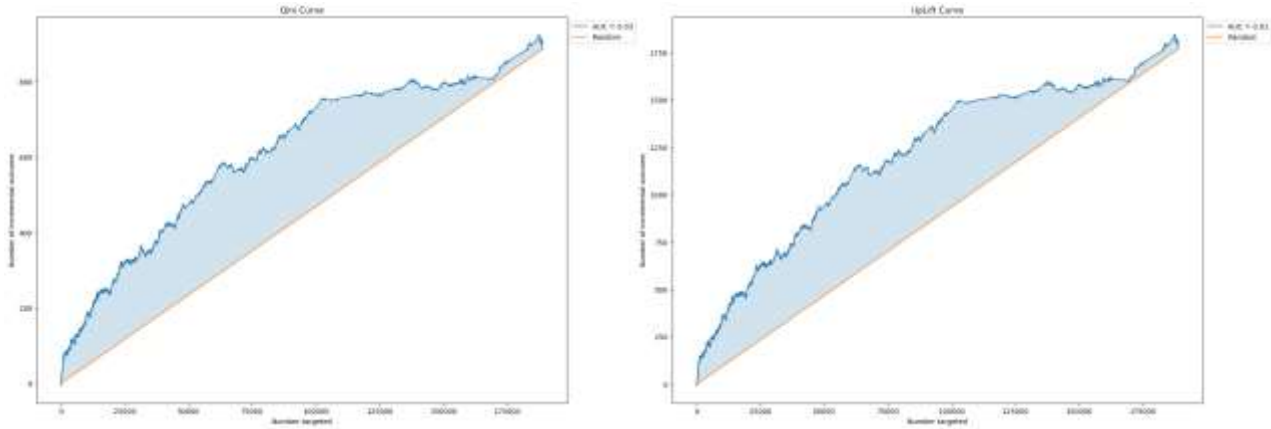

Рисунок 2.13 - Графики кривой QINI и UpLift для результатов моделирования с трансформацией класса с переходом к задаче регрессии в худшем случае

2.8 Результаты численного эксперимента

Проведя череду экспериментов, стоит посмотреть на все результаты разом и выделить лучшее решение для данных X5-Retail (рисунок 2.14).

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

Рисунок 2.14 - Сравнительные результаты целевых показателей качества обучения

Проведя череду экспериментов, стоит посмотреть на все результаты разом и выделить лучшее решение для собственных данных (рисунок 2.15).

Номер	Структура	$UpLift_{30\%}$	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

Рисунок 2.15 - Сравнительные результаты целевых показателей качества обучения

Стоит заметить, что в зависимости от данных, при одних и тех же подходах машинного обучения, наилучший результат дают совершенно разные модели

Как можно заметить, для наших данных по всем показателям (рисунок 2.15) лучшая модель для наших данных – это метод моделирования с помощью одной модели – стека из ансамблей моделей классификации под номером 6.

Далее найдем экономическую выгоду нашей модели с помощью показателя $UpLift_{30\%}$, т.к. он отражает номинальный прирост доли клиентов с покупкой вы выборке реципиентов. Пусть в среднем, клиент,

совершивший покупку, принесет 2 500 руб. выручки.

Изначально в нашем эксперименте участвовало 473 861 клиентов с отправкой СМС, что естественно не весь объем имеющейся базы и даже не 10% от нее. Тогда представим, что это 30% от имеющей базы для простоты интерпретации.

Из этих 473 тыс. реципиентов, покупку совершило 34 тыс., т.е. вероятность покупки примерно 0.0718 вне зависимости от объема выборки (при ее уменьшении). Наша наилучшая модель дает прирост в 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

ЗАКЛЮЧЕНИЕ

В данной выпускной квалификационной работе магистра предлагается исследование подходов к UpLift моделированию методами машинного обучения на исходных данных ретейл компании в сфере косметики и парфюмерии.

Были выбраны и описаны структуры с одной моделью машинного обучения, с двумя независимыми моделями машинного обучения и два вида трансформации класса для обучения одной модели машинного обучения классификации и регрессии.

Численные результаты эксперимента показали, что наилучшего UpLift по показателям качества обучения можно добиться с помощью автоматического подбора моделей задачи классификации и последующим применением ее в алгоритме с одной независимой моделью.

Найденный алгоритм, возможно, будет наилучшим только для рассматриваемых в задаче данных, так как в зависимости от скрытой природы зависимостей обучающих признаков, различные структуры могут показывать наилучшие результаты на одних данных и наихудшие на других.

В работе приведены обзоры на различные способы решения проблемы и полученные результаты в перспективе могут быть аналогичны и для остальной клиентской базы ретейл компании косметики и парфюмерии.

ЛИТЕРАТУРА

[1] Gutierrez P., G´erardy J. Causal Inference and Uplift Modeling A review of the literature // PMLR – 2016 - URL: https://proceedings.mlr.press/v67/gutierrez17a/gutierrez17a.pdf

[2] Weijia Zhang, Jiuyong Li, Lin Liu A unified survey of treatment effect heterogeneity modelling and uplift modelling // arXiv – 2021 – URL: https://arxiv.org/pdf/2007.12769

[3] Devriendt F., Guns T., Verbeke W. LEARNING TO RANK FOR UPLIFT MODELING // arXiv – 2020 – URL: https://arxiv.org/pdf/2002.05897

[4] Nyberg O., Kussmierczyk T., Klami A. Uplift Modeling with High Class Imbalance // PMLR – 2021 – URL: https://proceedings.mlr.press/v157/nyberg21a/nyberg21a.pdf

[5] RF – сегментация – URL: https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/

[6] Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение // пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс – 2018. – 652