

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»**

Журнал практики

Институт № 8 «Компьютерные науки и прикладная математика»

Кафедра 804 Учебная группа М8О-201М-21

ФИО обучающегося Фейзуллин Кирилл Маратович

Направление подготовки/
специальность 01.04.04 «Прикладная математика»

шифр, наименование направления подготовки/специальности

Вид практики Преддипломная

учебная, производственная, преддипломная или другой вид практики

Оценка за практику Платонов Е. Н.

1. Место и сроки проведения практики:

Наименование организации: МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ) кафедра 804

Сроки проведения практики

дата начала практики: 09 февраля 2023 г.

дата окончания практики: 10 мая 2023 г.

2. Инструктаж по технике безопасности:

подпись проводившего

/ Платонов Е.Н. /
расшифровка подписи

09 февраля 2023 г.
дата проведения

3. Индивидуальное задание обучающегося:

Разработать модель Decision trees for uplift modeling. Смоделировать обучающую выборку на основе данных для UpLift моделирования от ретейл компании в сфере косметики и парфюмерии.

4. План выполнения индивидуального задания обучающегося:

№ п/п	Место проведения	Тема	Период выполнения
1	Каф. 804	Инструктаж.	09.02.2023
2	Каф. 804	Знакомство с научной темой	09.02.23-28.02.23
3	Каф. 804	Работа с библиографическими источниками и ресурсами Интернет по теме исследования	01.03.23-31.03.23
4	Каф. 804	Разработка модели Decision trees for uplift modeling.	01.04.23-01.05.23
5	Каф. 804	Оформление отчета. Подведение итогов.	01.05.23-10.05.23

Утверждаю

подпись руководителя от МАИ

/ Платонов Е.Н. /
расшифровка подписи

09 февраля 2023 г.
дата утверждения*

подпись руководителя от
организации/предприятия

/ _____ /
расшифровка подписи

09 февраля 2023 г.
дата утверждения*

Ознакомлен

подпись обучающегося

/ Фейзуллин К.М. /
расшифровка подписи

09 февраля 2023 г.
дата ознакомления*

1. Отзыв руководителя практики от организации/предприятия:

Фейзуллину К.М. была поставлена задача исследования и разработки методов UpLift моделирования для рекламной кампании. Во время исследования были разобраны различные подходы для моделирования прогнозируемого инкрементального отклика реципиентов от рассылки сообщений. Для исследования были реализованы методы моделирования с помощью одной модели, с помощью двух независимых моделей и с помощью регрессионной модели, на языке программирования Python,

Выявлен лучший метод в виде одной модели классификации и выявлена лучшая модель классификации – стек из ансамбля моделей классификации.

Поставленная задача решена в полном объёме и заслуживает оценки «отлично».

_____/_____/____ 20__ г.
подпись руководителя от расшифровка подписи дата
организации/предприятия

2. Отчет обучающегося по практике:

Объектом исследования являются задача прогнозирования инкрементального отклика реципиентов рекламной кампании.

Цель работы – улучшение результатов работы UpLift моделирования методами машинного обучения.

Определение метрик для оценки качества UpLift моделирования

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.

UpLift на первых k – процентах выборки

Самая простая и интуитивно понятная метрика, особенно для применения в бизнесе и для интерпретации.

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.

Формула имеет следующий вид:

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

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

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 $k = 30\%$.

UpLift кривая (UpLift Curve)

Данная кривая строится как функция с нарастающим итогом, где для каждой точки задается соответствующий UpLift.

Определяется следующим образом:

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

$N_{target,Y=0,1}(t)$ – 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 .

Аналогично и для контрольной группы.

Пример данной кривой на рисунке 2.

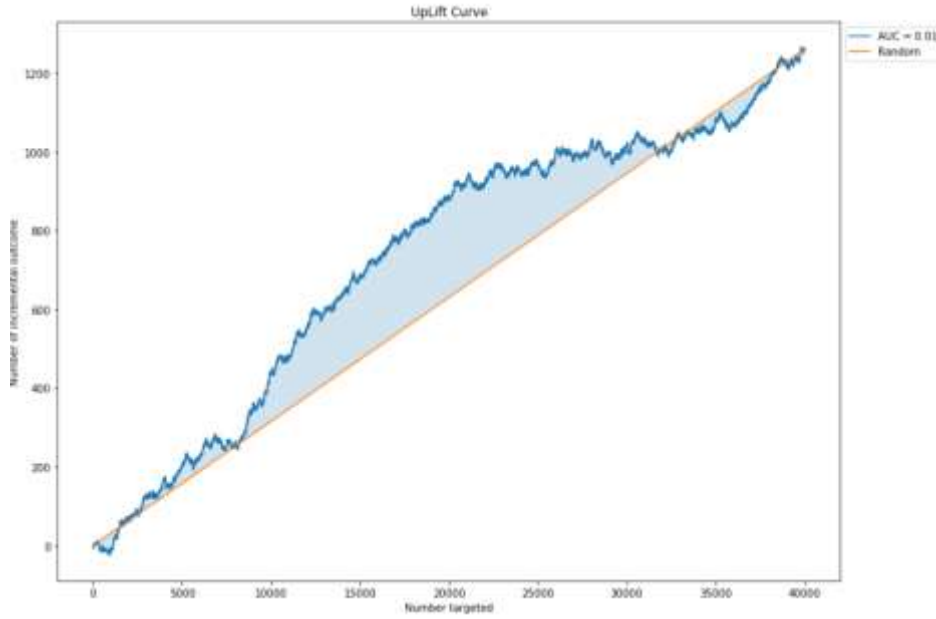


Рисунок 1

Qini кривая

Данную функцию можно выразить через UpLift кривую следующим образом:

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

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 производство 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.

Источник данных

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:

	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 2

- История покупок клиентов до коммуникаций:

	Карта лояльности	Дата покупки	Магазин покупки	Касса покупки	Чек покупки	Номенклатура	Сумма	ШТ. товара	Списано бонусов
1	0x4B0EADB857E761E6C4EF48775BC18F94	2022-10-31	AC5	AC6	100076050	CLOP32019	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	LOTLP002	602	1	34
5	0xE96996843A03020D3CC8D5A26A74BE8B	2022-10-31	AC5	AC6	100076062	SOD121304	473	1	26
6	0x3A8E58491A38FD31E3ABDEE59E60892E	2022-10-31	AC5	AC6	100076052	CLOP50067	169	1	0
7	0x4528A3C31F85ACF167A1AA6CA6F01D6	2022-10-31	AC5	AC6	100076053	POI759358	101	1	0
8	0x4528A3C31F85ACF167A1AA6CA6F01D6	2022-10-31	AC5	AC6	100076053	CLOP20097	349	1	0
9	0x44D8D5E9EBFD96D021D4A02D83B5C897	2022-10-31	AC5	AC6	100076058	CLOP31041	249	1	0
10	0x2F5AC08C159462C583729770CF0E93C7	2022-10-31	AC5	AC6	100076059	ELOR56120	249	1	0

Рисунок 3

Анализ и агрегирование данных

Так как данные для UpLift моделирования составляют 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.

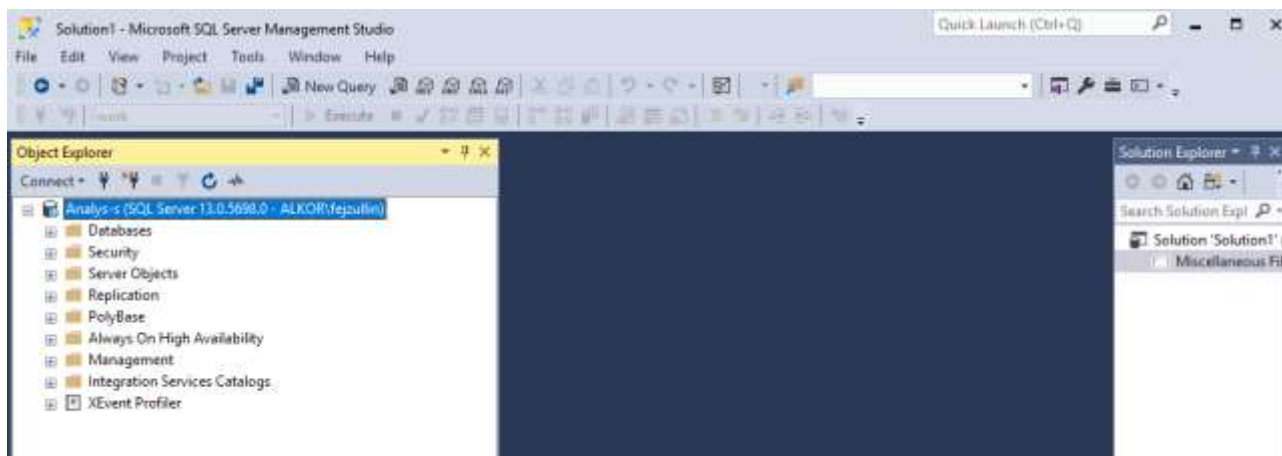


Рисунок 4

Для моделирования основных обучающих признаков был использован принцип RFM - сегментации¹. 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, минимальному 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

¹ RFM – сегментация // <https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/>

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:

Results	Messages	AMOUNT	ORDERS	AOV	last_order_days	LTV_W	AO_per_month	GET_Bonus	USE_Bonus	NON_SPEND_BONUS	BASE_GET_Bonus	BASE_USE_Bonus	Spends_GET_Bonus	Spends_USE_Bonus	Camp_GET_Bonus	POST_GET_Bonus	Camp_BASE_G
1	6a5b711b2c42	99.900000	1	99.900000	49	34.750000	0.250000000000000	207.000000	NULL	NULL	1.000000	NULL	300.000000	NULL	NULL	1.000000	NULL
2	6a5b711b2c42	NULL	0	NULL	NULL	NULL	0.000000000000000	1.000.000000	NULL	NULL	NULL	NULL	1.000.000000	NULL	NULL	NULL	NULL
3	6a5b711b2c42	NULL	0	NULL	NULL	NULL	0.000000000000000	9300.000000	NULL	NULL	NULL	NULL	1.000.000000	NULL	NULL	NULL	NULL
4	6a5b711b2c42	NULL	0	NULL	NULL	NULL	0.000000000000000	3000.000000	NULL	NULL	NULL	NULL	3000.000000	NULL	NULL	NULL	NULL
5	6a5b711b2c42	1035.200000	1	1035.200000	38	423.800000	0.250000000000000	318.000000	NULL	NULL	18.000000	NULL	300.000000	NULL	NULL	18.000000	NULL
6	6a5b711b2c42	247.000000	1	247.000000	94	81.760000	0.250000000000000	903.000000	247.000000	244.000000	0.000000	NULL	806.000000	247.000000	NULL	3.000000	NULL
7	6a5b711b2c42	1062.700000	1	1062.700000	14	263.187500	0.250000000000000	111.000000	NULL	NULL	11.000000	NULL	808.000000	NULL	NULL	11.000000	NULL
8	6a5b711b2c42	1418.050000	5	283.730000	38	264.862500	1.250000000000000	1078.000000	NULL	NULL	18.000000	NULL	1.000.000000	NULL	NULL	18.000000	NULL
9	6a5b711b2c42	3074.000000	1	3074.000000	29	760.500000	0.250000000000000	221.000000	300.000000	31.000000	31.000000	NULL	300.000000	300.000000	NULL	31.000000	NULL
10	6a5b711b2c42	1602.000000	3	534.000000	12	460.000000	0.750000000000000	817.000000	15.000000	602.000000	17.000000	15.000000	606.000000	NULL	NULL	17.000000	NULL
11	6a5b711b2c42	3648.200000	1	3648.200000	93	889.262500	0.250000000000000	1.037.000000	889.000000	817.000000	37.000000	NULL	1.000.000000	800.000000	NULL	37.000000	NULL

Рисунок 5

Реализация UpLift моделирования методами машинного обучения

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 получаем следующие значения метрик:

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

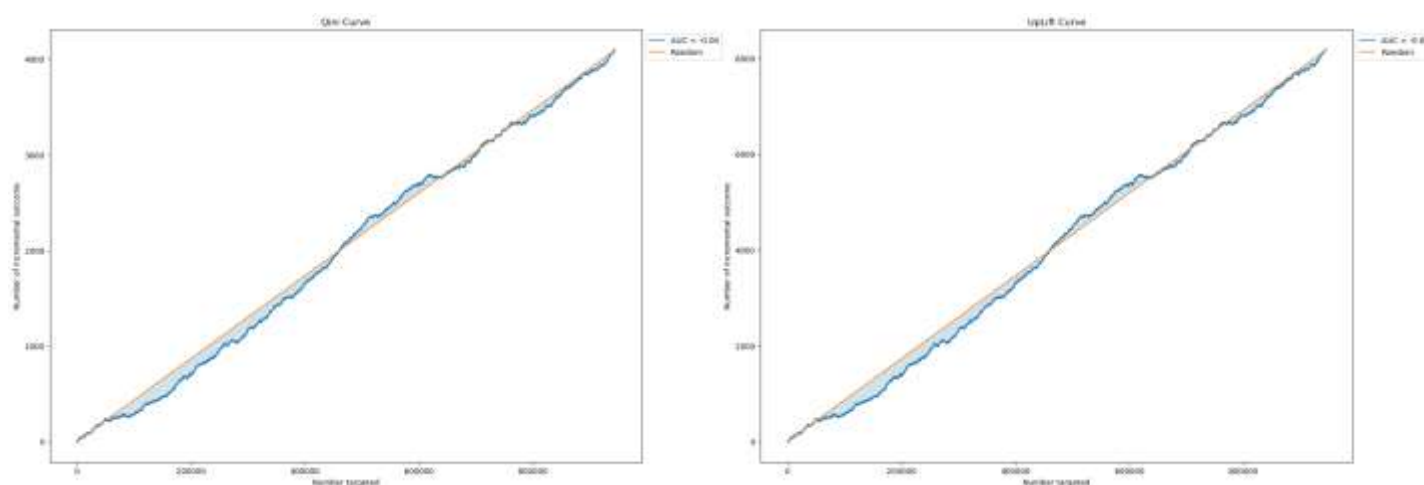


Рисунок 6. Графики кривой QINI и UpLift

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 с разбиением выборки на 5 долей. По итогу кросс валидации будет браться средняя по метрикам качества, на основе которых и будет сравнение. Иллюстрация работы кросс валидации на рисунке 13.

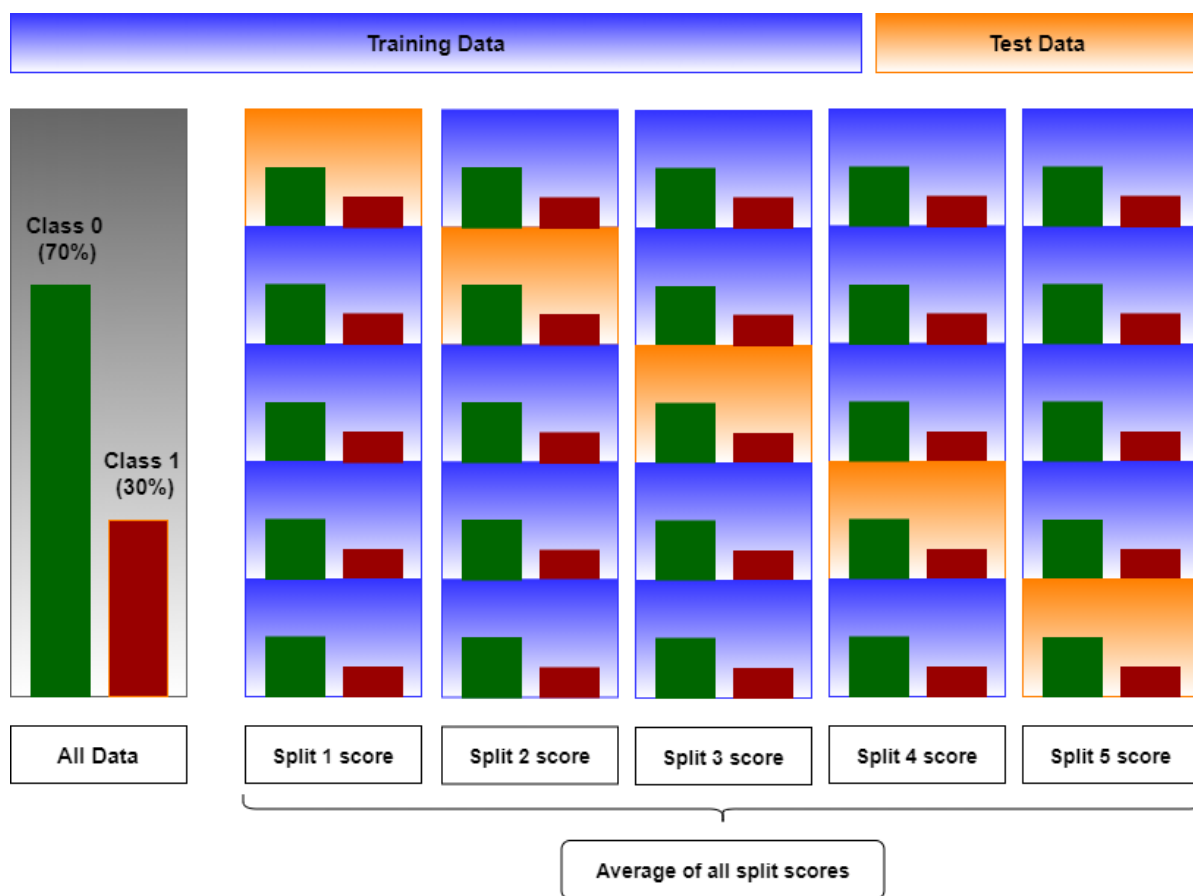


Рисунок 7. Схема кросс валидации

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 моделирования получены следующие усредненные метрики:

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

По итогу кросс валидации имеются два типа событий:

- Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 8.

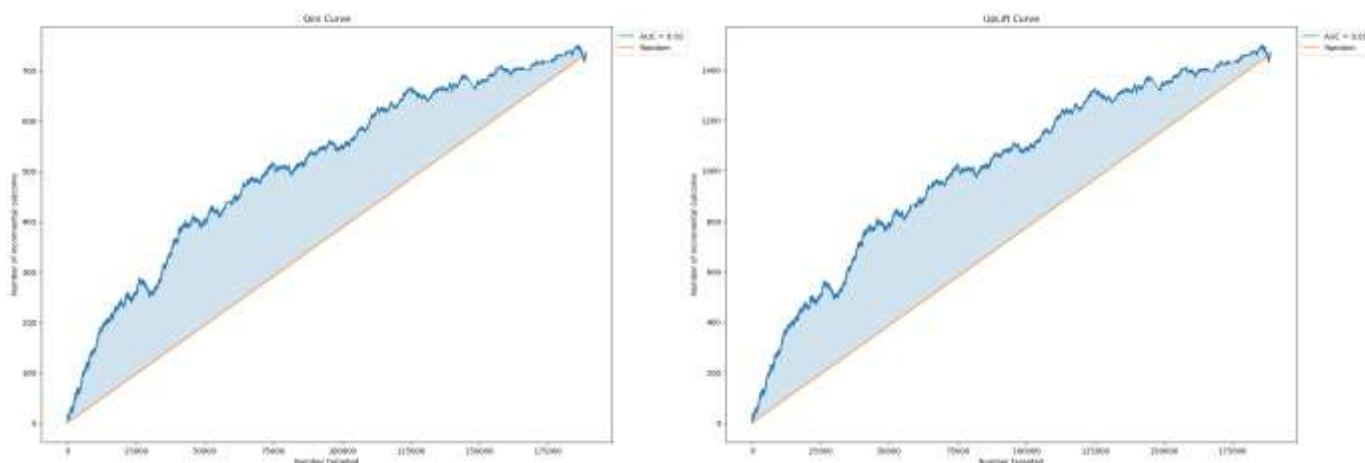


Рисунок 8. Графики кривой QINI и UpLift для результатов моделирования одной моделью в лучшем случае

- Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 9.

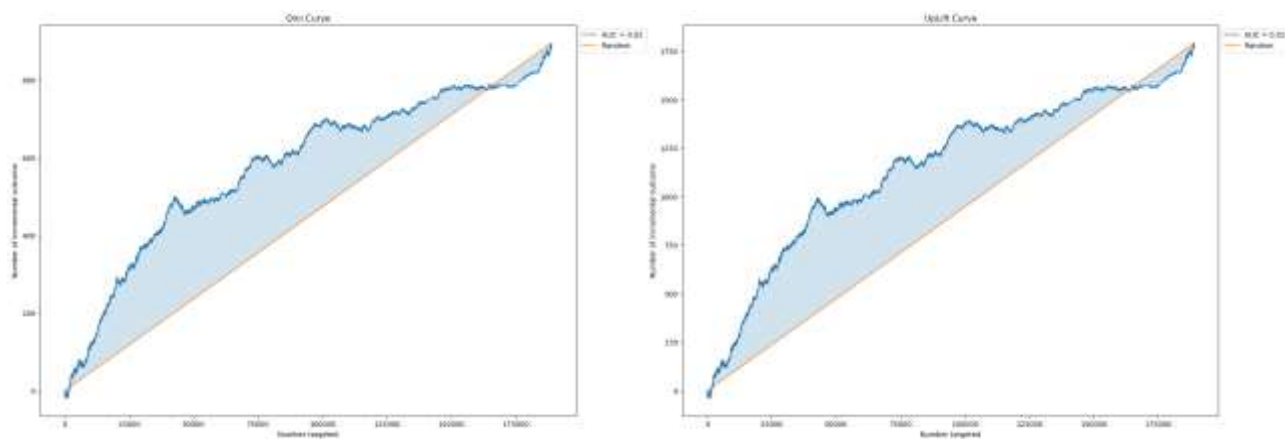


Рисунок 9. Графики кривой QINI и UpLift для результатов моделирования одной моделью в худшем случае

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.

По итогам моделирования получены следующие усредненные метрики:

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

По итогу кросс валидации имеются два типа событий:

- Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 10.

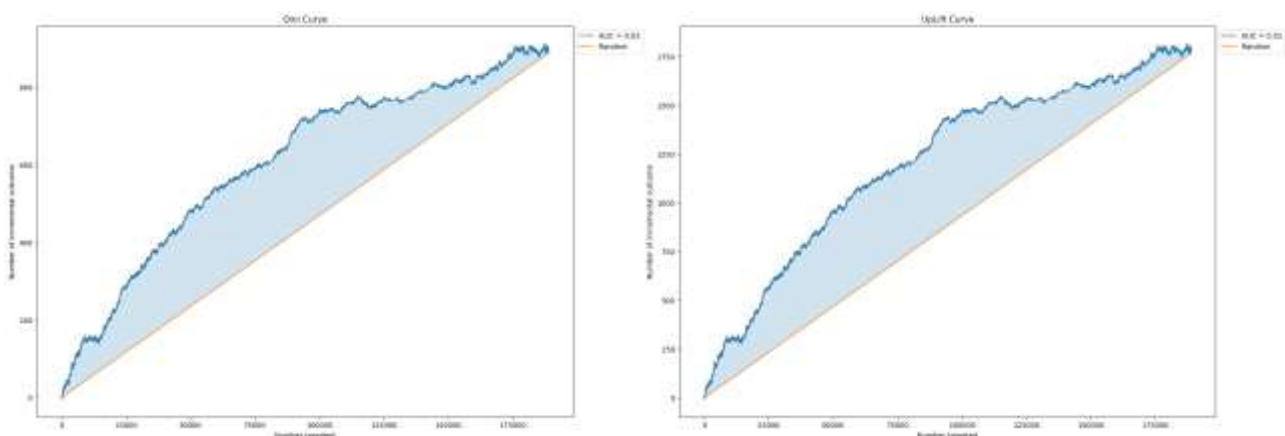


Рисунок 10. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

- Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 11.

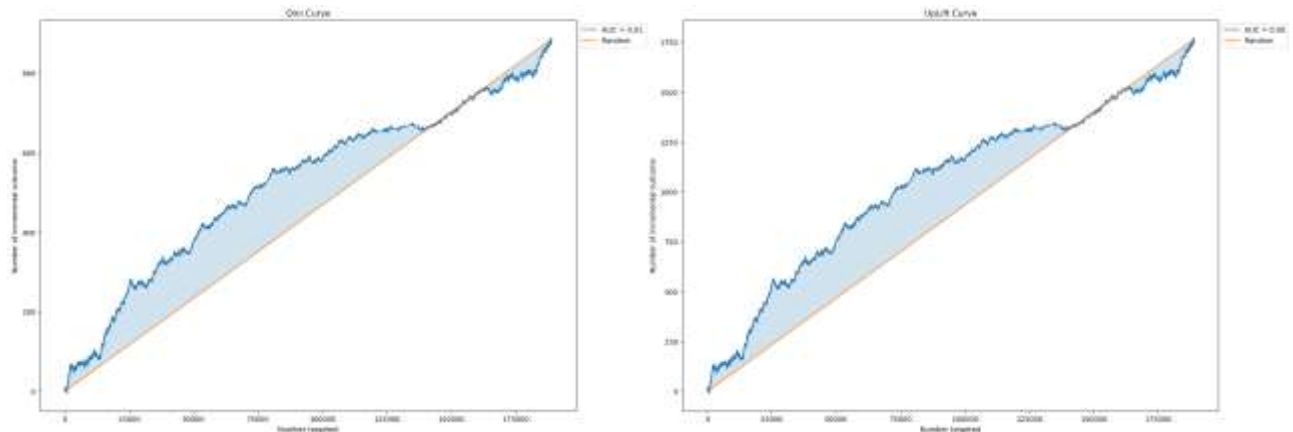


Рисунок 11. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

Так же стоит добавить, что поведение показателей качества обучения на тестовой выборке в 4 из 5 итераций кросс валидации выглядит как на рисунке 11, что говорит об ухудшении качества обучения – о чем и сигнализируют усредненные показатели $UpLift_{30\%}$, Qini curve AUC, UpLift curve AUC.

Метод трансформации класса (задача классификации)

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

$Z_i = Y_i * W_i + (1 - Y_i)(1 - W_i)$, где Y_i – целевая переменная, W_i – коммуникационная переменная.

Тогда трансформированный класс будет иметь следующие значения:

$$Z_i = \begin{cases} 1 & \text{при } W_i = 1; Y_i = 1 \\ 0 & \text{при } W_i = 0; Y_i = 1 \\ 0 & \text{при } W_i = 1; Y_i = 0 \\ 1 & \text{при } W_i = 0; Y_i = 0 \end{cases}$$

Далее произведем переход к задаче классификации для однозначной интерпретации прогноза.

По результатам моделирования были получены следующие усредненные результаты:

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

По итогу кросс валидации имеются два типа событий:

- Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 12.

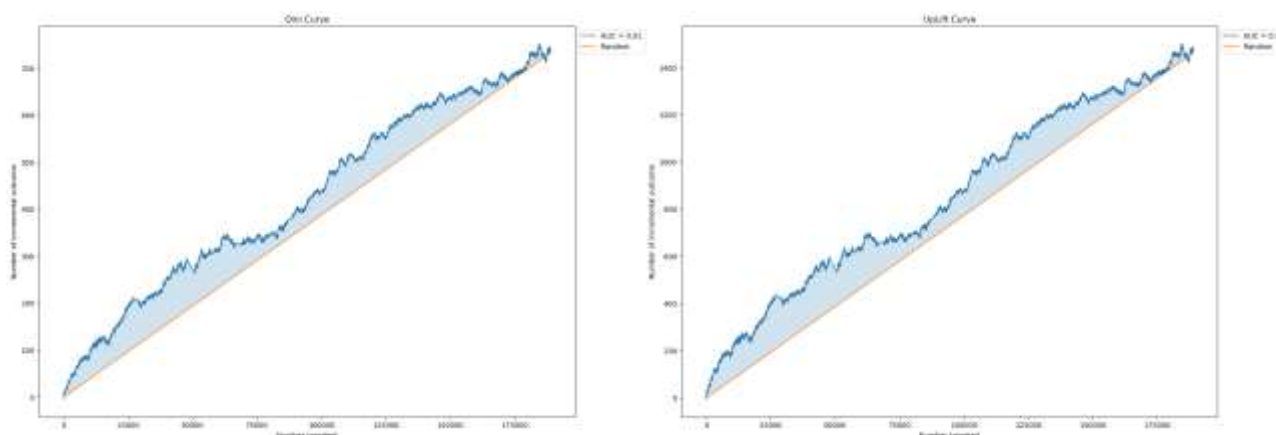


Рисунок 12. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

- Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 13.

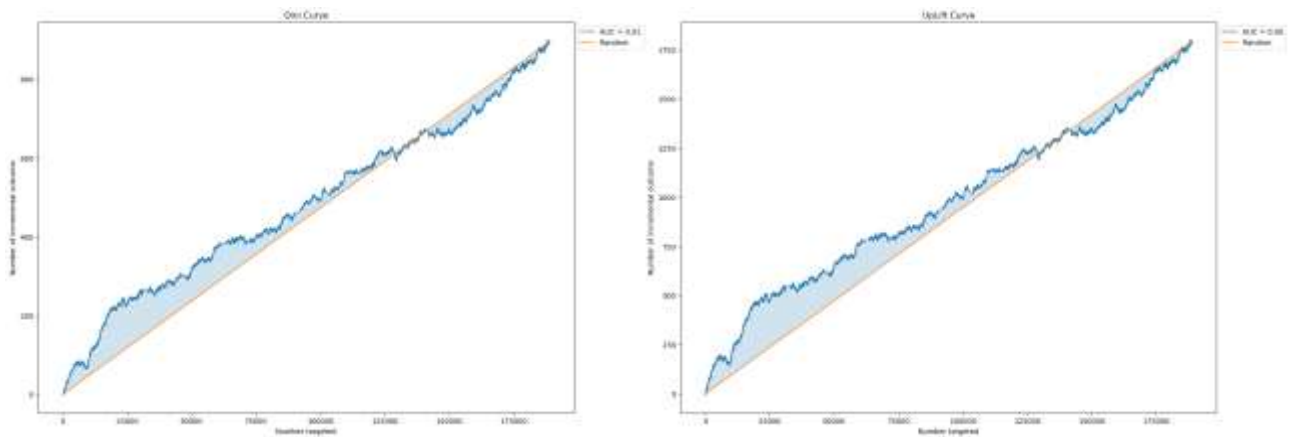


Рисунок 13. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

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

Метод трансформации класса (задача регрессии)

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

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

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

$$Z_i = \begin{cases} 2, & \text{при } W_i = 1; Y_i = 1 \\ 0, & \text{при } W_i = 0, 1; Y_i = 1 \\ -2, & \text{при } W_i = 0; Y_i = 1 \end{cases}$$

Далее произведем переход к задаче регрессии для однозначной интерпретации прогноза.

По результатам моделирования были получены следующие усредненные результаты:

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

По итогу кросс валидации имеются два типа событий:

- Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 14.

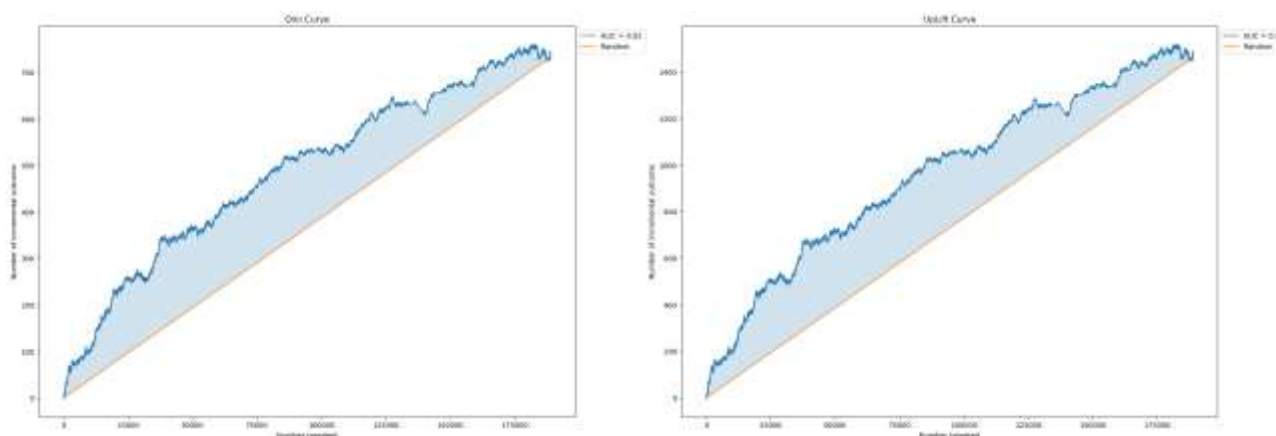


Рисунок 14. Графики кривой QINI и Uplift для результатов моделирования с двумя моделями в лучшем случае

- Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 15.

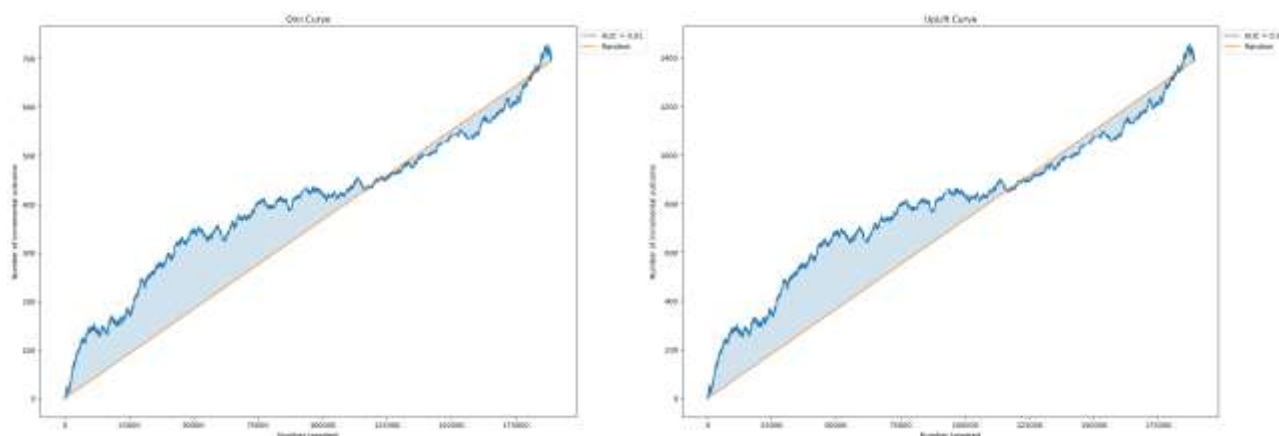


Рисунок 15. Графики кривой QINI и Uplift для результатов моделирования с двумя моделями в худшем случае

Данный метод чуть хуже, чем метод с двумя независимыми моделями.

Исследований архитектур моделей машинного обучения

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 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 1.

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%

Таблица 1

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

По результатам моделирования были получены следующие усредненные результаты:

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

По итогу кросс валидации имеются два типа событий:

- Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 16.

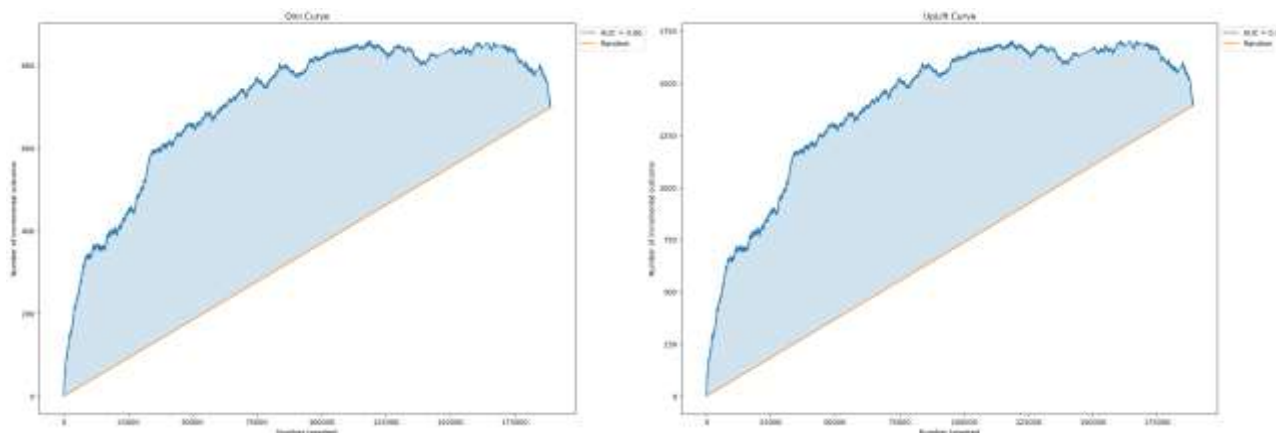


Рисунок 16. Графики кривой QINI и Uplift для результатов моделирования с двумя моделями в лучшем случае

- Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 17.

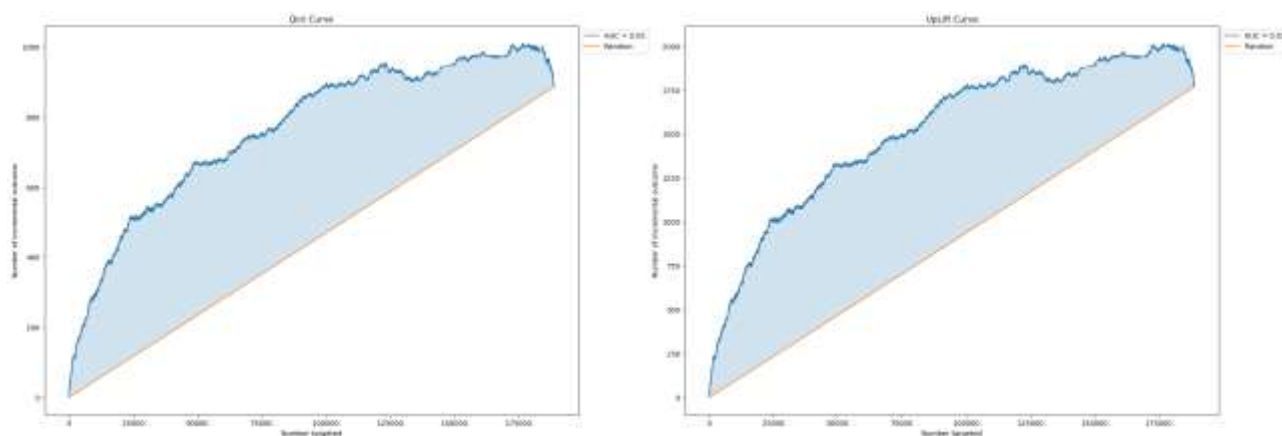


Рисунок 17. Графики кривой QINI и Uplift для результатов моделирования с двумя моделями в худшем случае

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.

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%

Таблица 2.

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

По результатам моделирования были получены следующие усредненные результаты:

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

По итогу кросс валидации имеются два типа событий:

- Когда uplift позволяет получить наибольший инкрементальный эффект, как на рисунке 18.

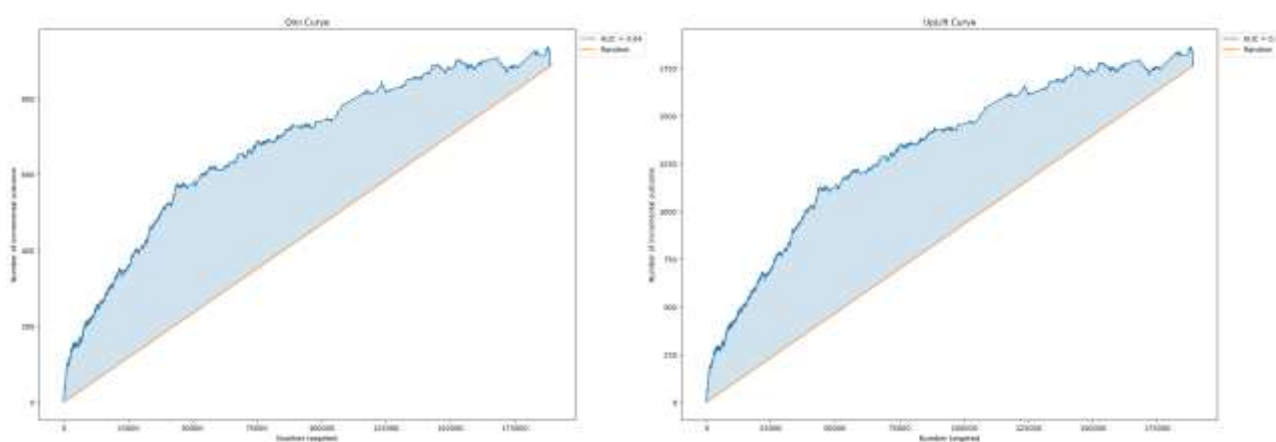


Рисунок 18. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в лучшем случае

- Когда uplift позволяет получить инкрементальный эффект с переменным успехом, как на рисунке 19.

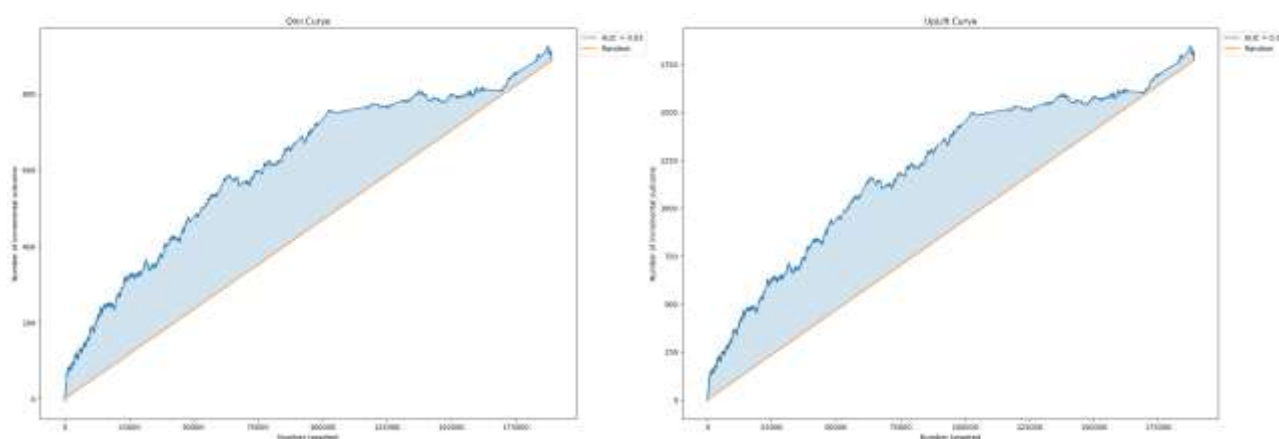


Рисунок 19. Графики кривой QINI и UpLift для результатов моделирования с двумя моделями в худшем случае

Сравнительные результаты

Проведя череду экспериментов, стоит посмотреть на все результаты разом и выделить лучшее решение – результаты на рисунке 20.

Номер	Структура	<i>UpLift</i> _{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 20. Сравнительные результаты целевых показателей качества обучения

Как можно заметить, по всем показателям 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 работе были исследованы методы моделирования UpLift с помощью машинного обучения на исходных данных ретейл компании в сфере косметики и парфюмерии.

В работе были рассмотрены метрики оценивания качества прогноза UpLift при алгоритме с одной моделью, при алгоритме с двумя независимыми моделями и при работе с одной моделью после трансформации классов и перехода к задаче классификации и регрессии.

По итогам моделирования с данными обучающими признаками, это метод моделирования с помощью одной модели.

После определения метода было решено найти наилучшую структуру модели с помощью AutoML конвейеров. В результате чего выяснилось, что с данными признаками лучшей моделью является стека из ансамблей моделей классификации.

Причем, при использовании стека в методе с одной моделью, данный алгоритм имеет наилучшие показатели по всем целевым метрикам по итогам усреднения результатов кросс – валидации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. RF – сегментация // <https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/>

_____ / _____ 20__ г.
подпись обучающегося *расшифровка подписи* *дата*