МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

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

Производственная (Профессиональная практика)

Выполнил:
студент группы М8О-201М-21
Фейзуллин К.М.

Научный руководитель:
к.ф.-м.н., доцент
Платонов Е.Н.

Москва, 2022

Актуальность

- Актуальность данной работы появилась с ростом потребительской экономики
- Реализация данной задачи возможна благодаря массовой цифровизации

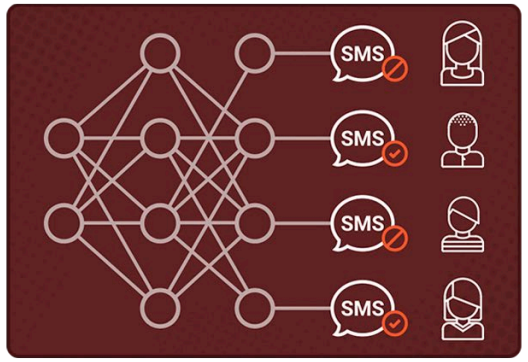
Цель работы

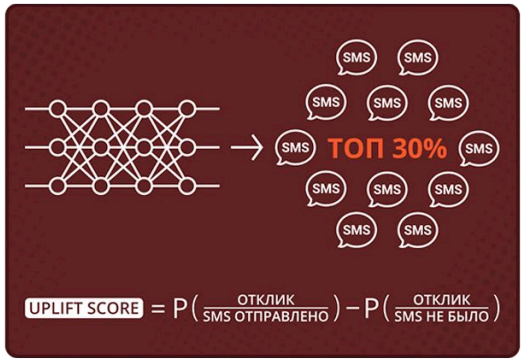
• Разработать модель Decision trees for uplift modeling. Смоделировать обучающую выборку на основе данных для UpLift моделирования от X5 Retail с добавлением характеристик заработка и списания бонусов.

Задачи:

- Описание набора данных;
- Выбор метрик качества;
- Выбор используемых моделей и исследование их качества работы;
- Сравнение полученных результатов.

Описание набора данных





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 15:40:48.0000000	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

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

Метрики качества - 1

- UpLift на первых k – процентах выборки:

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

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

- 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.).

Метрики качества - 2

- UpLift кривая (UpLift Curve):

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

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

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

Выбор используемых моделей и исследование их качества работы.

Модель	WAU	UpLift на k%	Qini AUC	UpLift AUC
Базовая модель	3,32%	3,41%	0,00%	0,00%
Одна предиктивная модель	3,32%	3,19%	0,00%	0,00%
Две независимые предиктивные модели	3,33%	5,34%	1,00%	1,20%
Регрессионная модель с трансформацией класса	3,29%	4,41%	0,60%	0,60%
Наилучшего найденный конвейра машинного обучения	3,47%	6,99%	2,40%	3,40%

Выводы

- Новые показатели начисления и списания бонусов, и их производные, благоприятно сказались на метриках моделей.

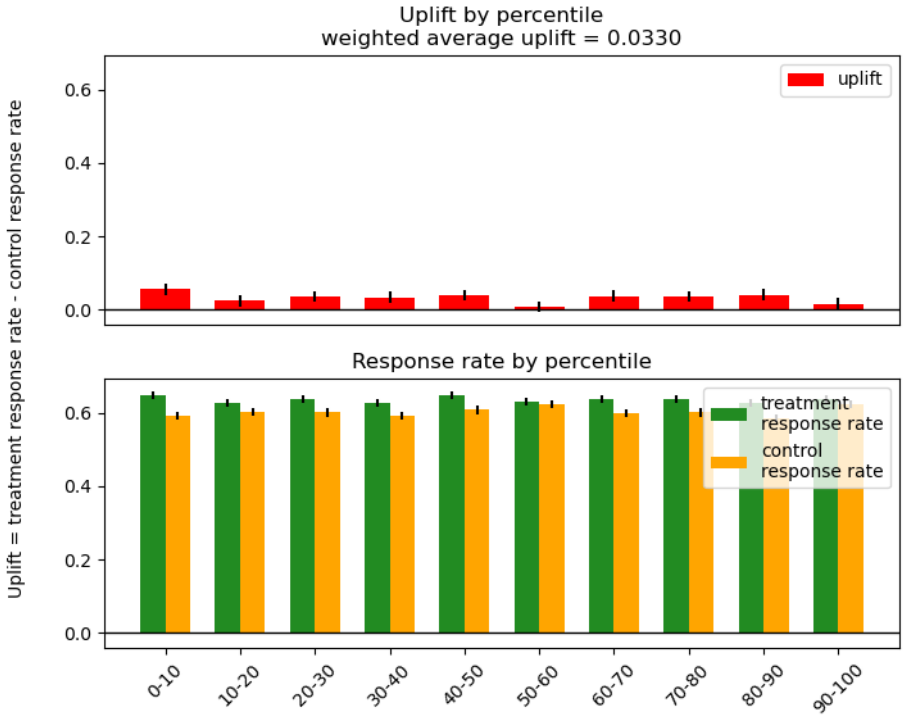
- Качество модели очень зависит от ее архитектуры и гипер-параметров.

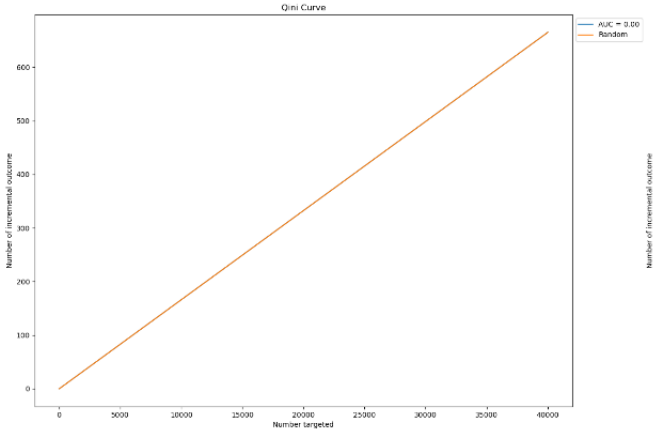
- Для дальнейшего улучшения модели нужно модернизировать автоматический поиск наилучшей архитектуры и оптимизации гипер-параметров под задачу UpLift

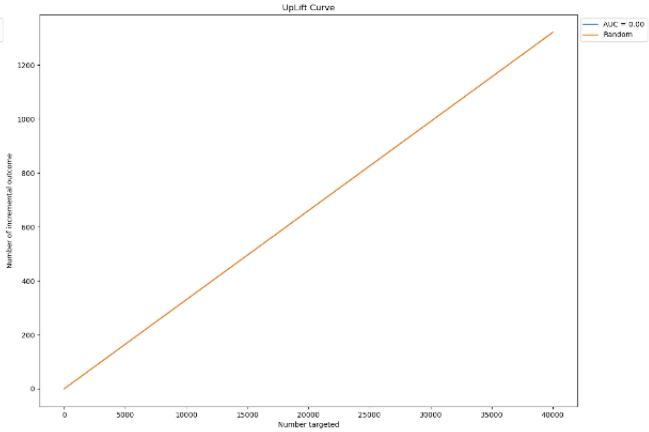
Спасибо за внимание !

Графические результаты работы моделей - 1

- Решение с одной моделью

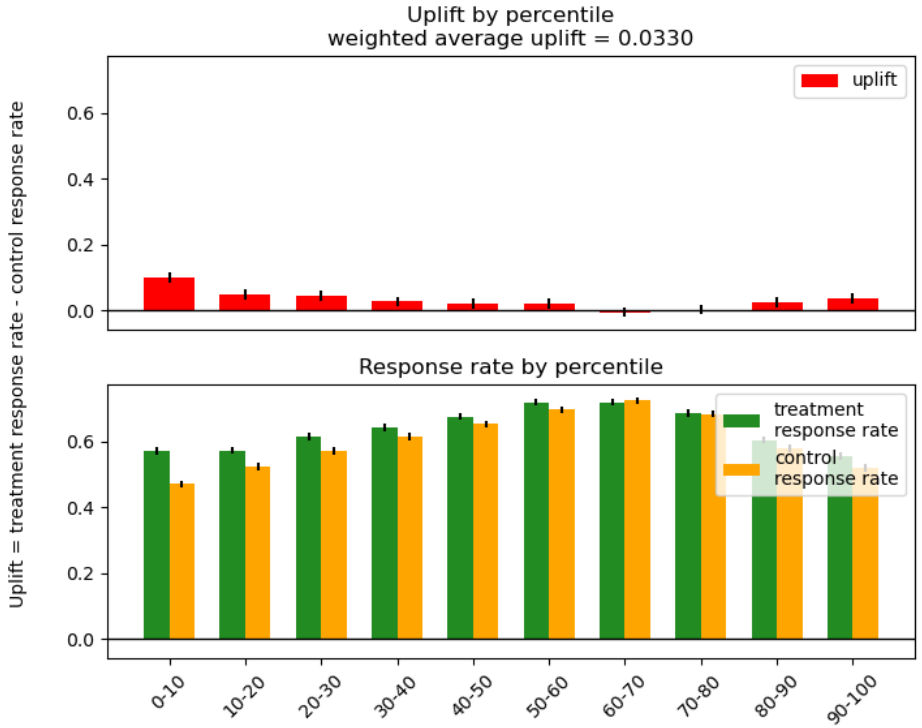


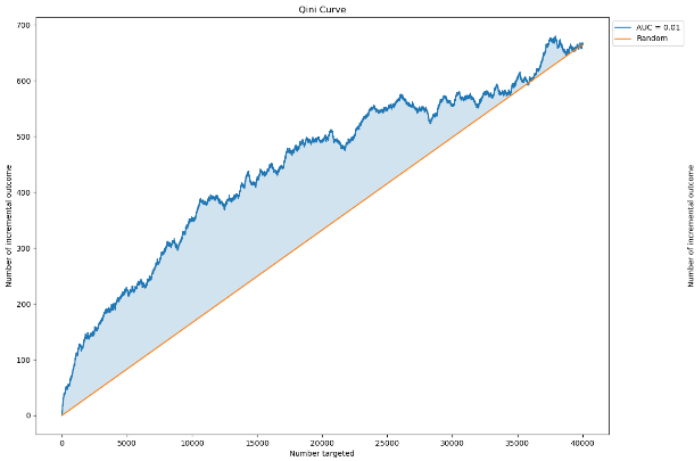


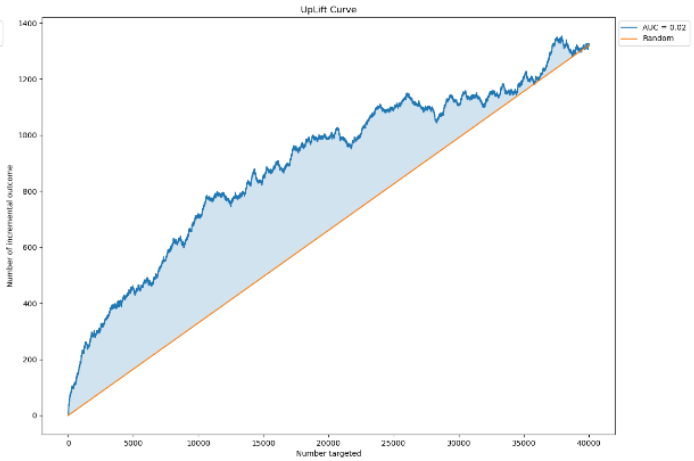


Графические результаты работы моделей - 2

- Решение с двумя независимыми моделями

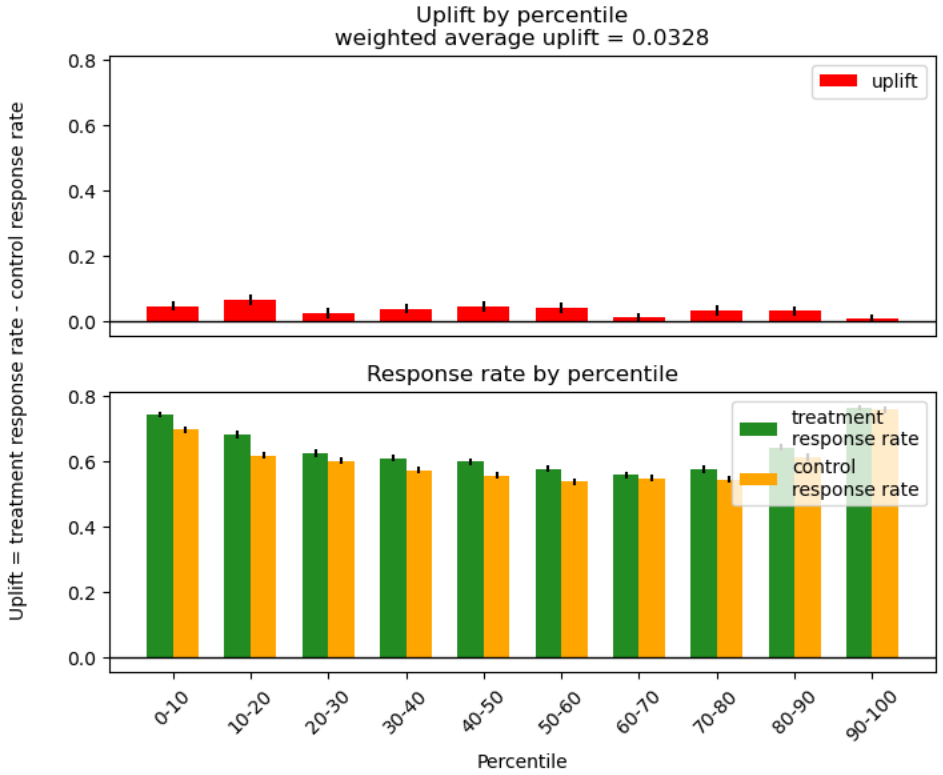


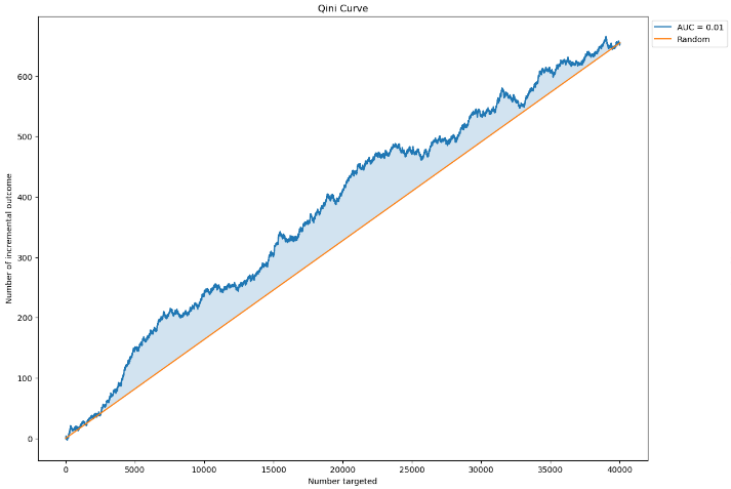


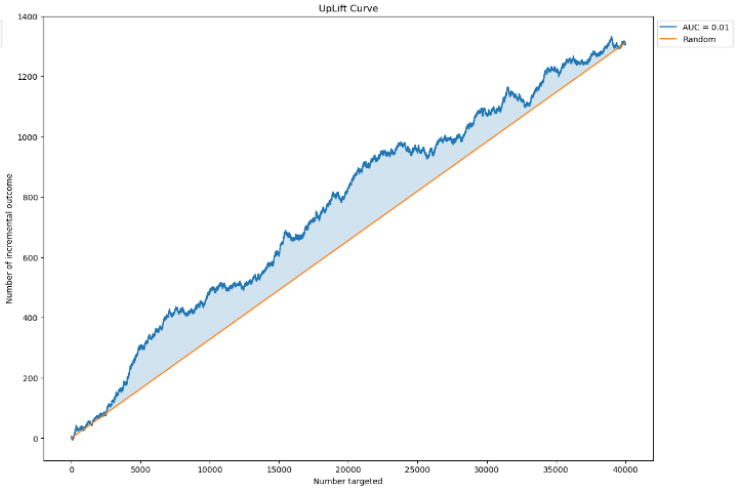


Графические результаты работы моделей - 3

- Трансформация класса с переходом к задаче регрессии







Графические результаты работы моделей - 4

- Лучший PipeLine