ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТИ ПОДКЛЮЧЕНИЯ КЛИЕНТОМ УСЛУГИ МОБИЛЬНОГО ОПЕРАТОРА

А.В. Дунаев

Постановка задачи

Задача

У нас появился запрос из отдела продаж и маркетинга. Оператор сотовой связи предлагает обширный набор различных услуг своим абонентам. При этом, разным пользователям интересны разные услуги. Поэтому, необходимо построить алгоритм, который для каждой пары пользователь-услуга определит вероятность подключения услуги.

Данные

В качестве исходных данных вам будет доступна информация об отклике абонентов на предложение подключения одной из услуг. Каждому пользователю может быть сделано несколько предложений в разное время, каждое из которых он может или принять, или отклонить.

Отдельным набором данных будет являться нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента. Эти данные привязаны к определенному времени, поскольку профиль абонента может меняться с течением времени.

Данные train и test разбиты по периодам – на train доступно 4 месяца, а на test отложен последующий месяц.

Итого, в качестве входных данных будут представлены:

- data_train.csv: id, vas_id, buy_time, target;
- features.csv.zip: id, <feature_list>.

И тестовый набор:

- data_test.csv: id, vas_id, buy_time;
- target - целевая переменная, где 1 означает подключение услуги, 0 - абонент не подключил услугу соответственно;
- buy_time - время покупки, представлено в формате timestamp, для работы с этим столбцом понадобится функция datetime.fromtimestamp из модуля datetime;
- id - идентификатор абонента;
- vas_id - подключаемая услуга.

Метрика

Скоринг будет осуществляться функцией f1, невзвешенным образом, как, например, делает функция sklearn.metrics.f1_score(…, average='macro').

Анализ признака ‘not_first_offer’ (не первое предложение)

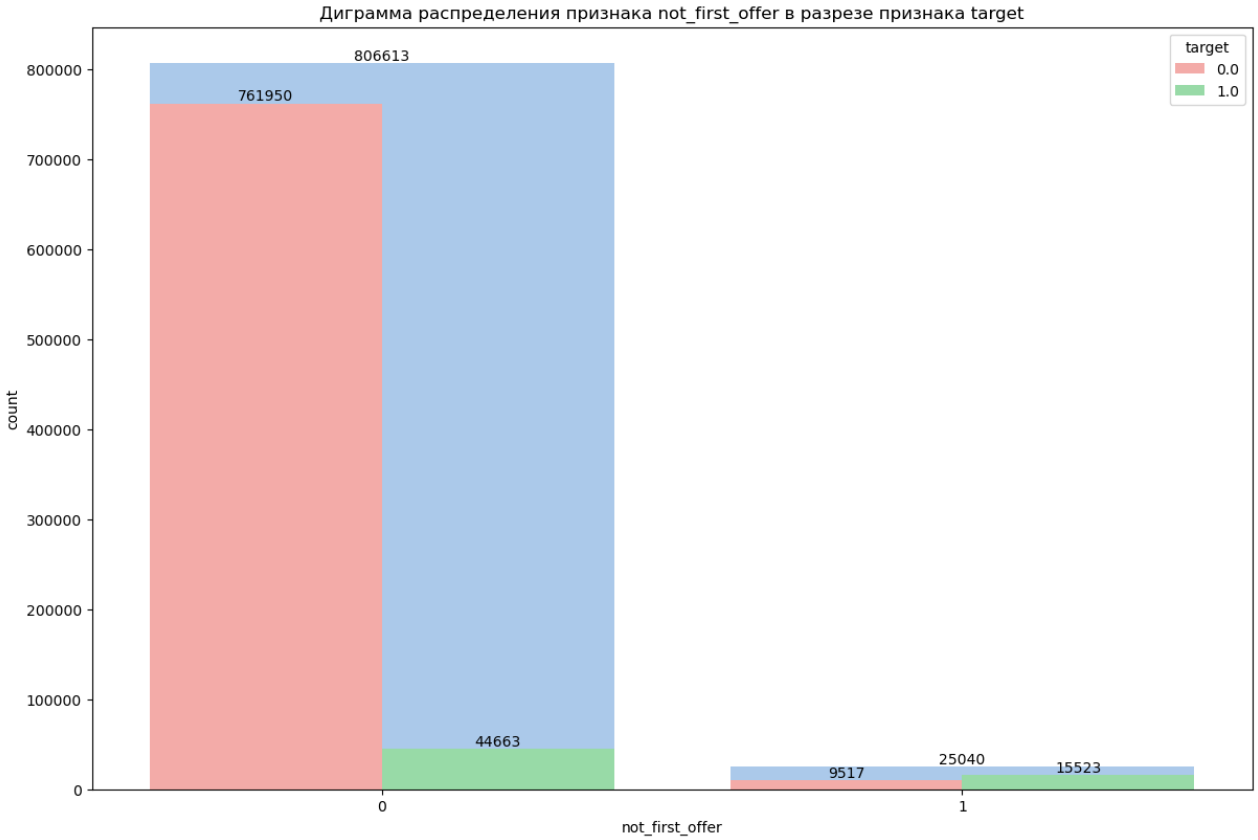


Таблица распределения частот

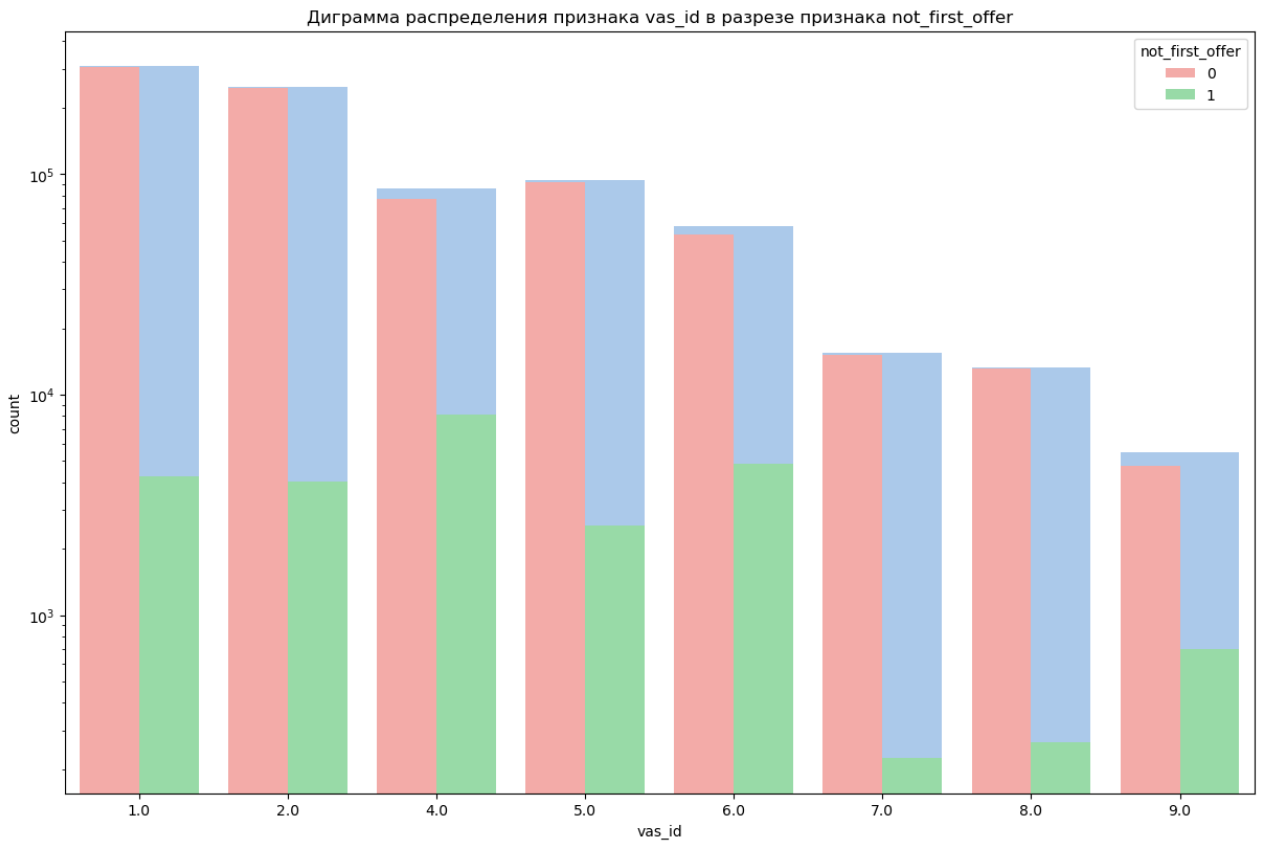
	target		sum
	0.0	1.0	
not_first_offer			
0	761950	44663	806613
1	9517	15523	25040
sum	771467	60186	831653

Таблица распределения относительных частот

	target		sum
target	0.0	1.0	
not_first_offer			
0	0.94	0.06	1
1	0.38	0.62	1

Как видно из представленных данных, доля абонентов, совершивших покупку при первом предложении, гораздо меньше, чем доля абонентов, которым было осуществлено большее количество предложений (относительные частоты 0.06 и 0.62 соответственно). При этом, общее количество повторных предложений сделано очень мало (всего 25040 при общем количестве предложений 831653). В качестве возможной рекомендации предлагается увеличить количество повторных предложений услуг.

Исследование связи признака 'vas_id’ (подключаемая услуга) с признаком 'not_first_offer’ (не первое предложение)



Как видно из представленных данных, повторное предложение услуг 7 и 8 осуществлялось гораздо реже остальных. В качестве возможной рекомендации предлагается для данных услуг в первую очередь увеличить количество повторных предложений.

Анализ признака ‘vas_id’ (подключаемая услуга)

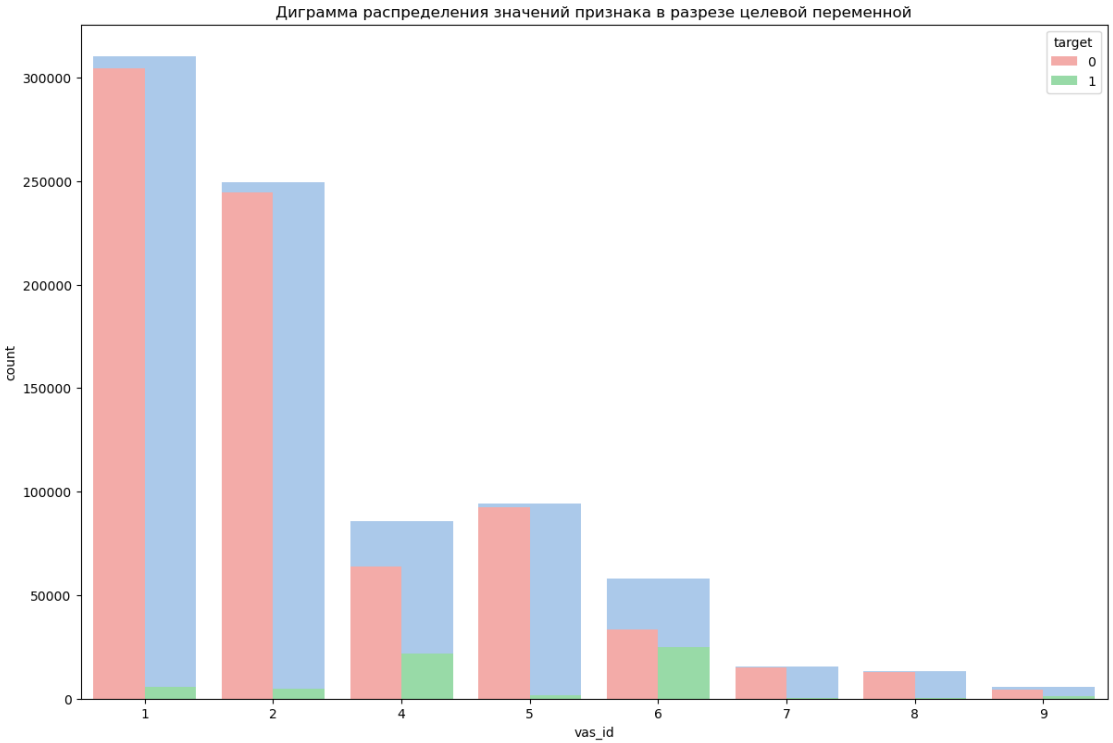


Таблица распределения частот

	target		sum
target	0.0	1.0	
vas_id			
1.0	304511	5664	310175
2.0	244708	4797	249505
4.0	63991	21765	85756
5.0	92393	1692	94085
6.0	33174	24704	57878
7.0	15219	213	15432
8.0	13003	347	13350
9.0	4468	1004	5472
sum	771467	60186	831653

Таблица распределения относительных частот

target			sum
target	0.0	1.0	
vas_id			
1.0	0.98	0.02	1
2.0	0.98	0.02	1
4.0	0.75	0.25	1
5.0	0.98	0.02	1
6.0	0.57	0.43	1
7.0	0.99	0.01	1
8.0	0.97	0.03	1
9.0	0.82	0.18	1

Как видно из представленных данных, услуги 4, 6 и 9 пользовались относительно большим успехом у абонентов, т.к. они чаще всего подключались (относительные частоты 0.25, 0.43 и 0.18 соответственно). При этом, в сравнении с услугами 4 и 6, услуга 9 гораздо реже предлагалась абонентам - всего 5472 раз, в то время как услуги 4 и 6 - 85756 и 57878 соответственно. Услуги 1 и 2 чаще остальных предлагались абонентам, однако, они не пользовались успехом (относительные частоты составили менее 0.02). В качестве возможной рекомендации предлагается увеличить количество предложений по услугам 4, 6 и 9, а по услугам 1 и 2 - повысить качество предложений.

Анализ признака ‘buy_month’ (месяц, в который была предложена услуга)

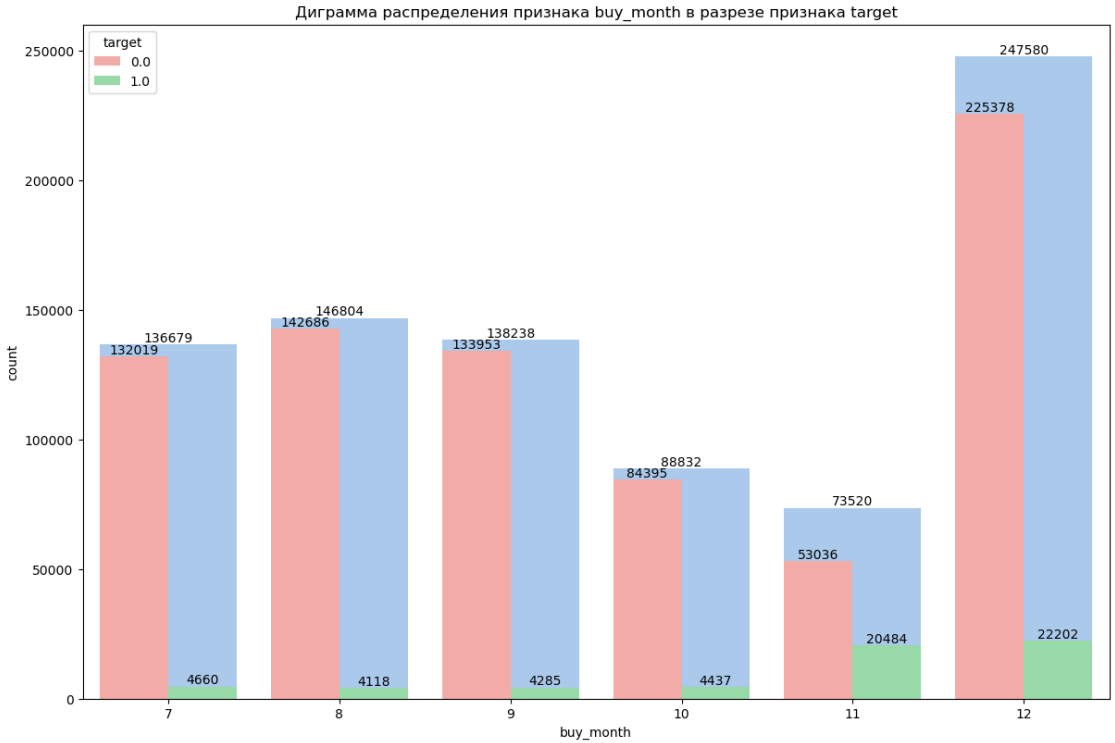


Таблица распределения частот

target		sum	
target	0.0	1.0	
buy_month			
7	132019	4660	136679
8	142686	4118	146804
9	133953	4285	138238
10	84395	4437	88832
11	53036	20484	73520
12	225378	22202	247580
sum	771467	60186	831653

Таблица распределения относительных частот

target		sum	
target	0.0	1.0	
buy_month			
7	0.97	0.03	1
8	0.97	0.03	1
9	0.97	0.03	1
10	0.95	0.05	1
11	0.72	0.28	1
12	0.91	0.09	1

Как видно из представленных данных, ноябрь ('buy_month' = 11) отличается гораздо большей относительной частотой подключения - 0.28, в то время, как относительные частоты подключения в другие месяцы не превышали 0.09. При этом, количество предложений в ноябре оказалось самым низким. Поскольку отсутствуют данные за другие годы, то не представляется возможность сделать выводы о наличии месячной закономерности. В качестве возможной рекомендации предлагается увеличить количество предложений в ноябре.

Анализ признака «target» (целевая переменная)

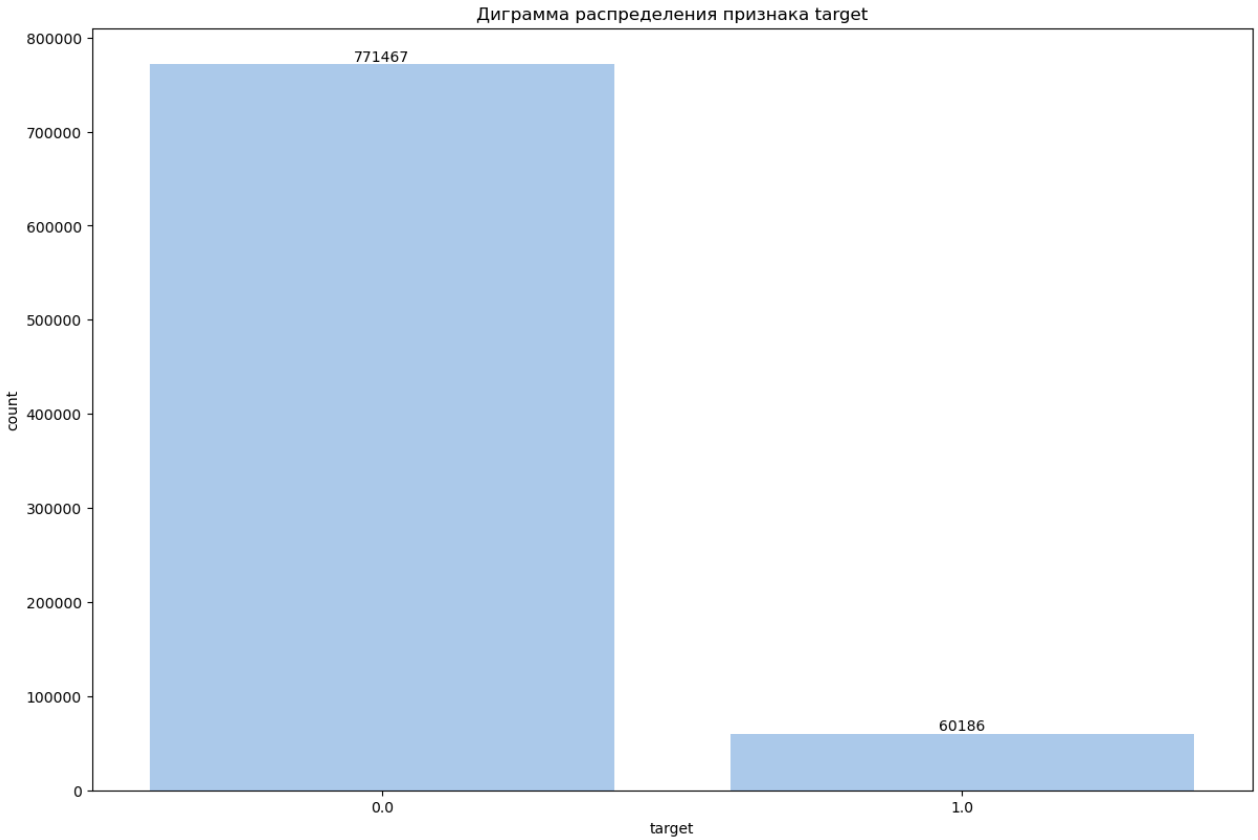


Таблица распределения частот

	target		sum
	0.0	1.0	
not_first_offer			
0	761950	44663	806613
1	9517	15523	25040
sum	771467	60186	831653

Таблица распределения относительных частот

	target		sum
target	0.0	1.0	
not_first_offer			
0	0.94	0.06	1
1	0.38	0.62	1

Целевая переменная 'target' является бинарно-количественным признаком с высокой несбалансированностью (количество значений 0 многократно привышает количество значений 1). Для обеспечения лучшей сбалансированности при валидации моделей предлагается использовать стратиграфированные валидационные выборки (например, метод «Repeated Stratified KFold»).

Отбор признаков и выбор модели

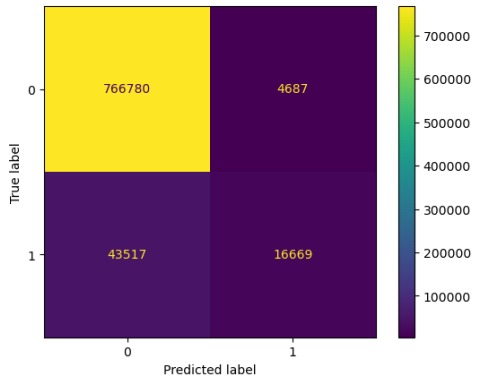
В процессе выбора модели были опробованы следующие варианты:

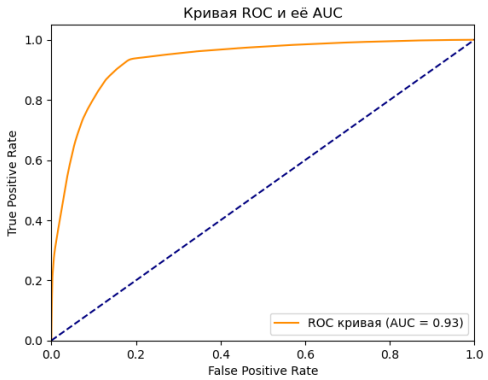
- «Logistic Regression»;
- «Decision Tree Classifier»;
- «XGBoost Classifier».

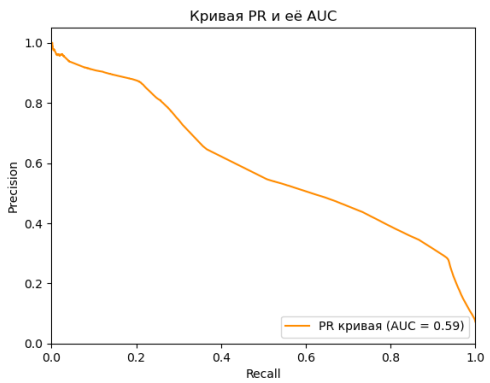
Отбор признаков осуществлялся для каждой из опробованной модели методом «Select From Model» с экспертной отсечкой признаков по уровню значимости. Лучший результат на валидационных данных, полученных методом «Repeated Stratified KFold» показала модель «XGBoost Classifier». Итоговая точность модели по метрике «f1 macro» на валидационной выборке составила 0.69.

Classification report

	precision	recall	f1-score	support
0	0.95	0.99	0.97	771467
1		0.28	0.41	60186
accuracy			0.94	831653
macro avg	0.86	0.64	0.69	831653
weighted avg	0.93	0.94	0.93	831653







Принцип составления индивидуальных предложений для выбранных абонентов

- увеличить количество повторных предложений услуг, в первую очередь, это относится к услугам 7 и 8;
- увеличить общее количество предложений по услугам 4, 6 и 9, а по услугам 1 и 2 - повысить качество предложений;
- увеличить общее количество предложений услуг в ноябре.