РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

цифровой прорыв

сезон: ИИ

Министерство экономического развития Российской Федерации

Цифровой прорыв 2023

Международный хакатон (Москва)

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

цифровой прорыв

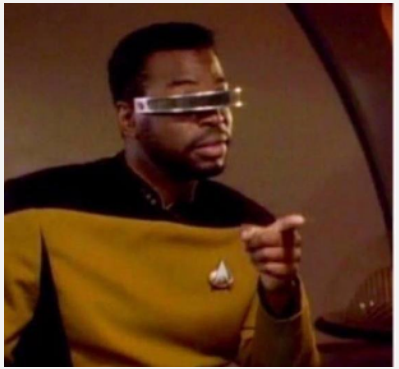
сезон: ИИ

Министерство экономического развития Российской Федерации

Каледин Артем

Modelling / Features

@htklf



Петров Артем

EDA / Tuning

Щипков Никита

EDA

Знакомьтесь, команда

BEE DONALDS !

Задание

- Необходимо построить предиктивный сервис, будет предсказывать вероятность того, что клиент не попадет в отток.

Выбор модели алгоритма

- Протестированы линейные модели и модель бустинга. Итоговым решением была выбрана модель CatBoost

F1-score на тестовой выборке (private) -> 0,641
RMSE (oot valid) ~14.5

Таргет-то не настоящий!

1. В каждой строчке – по одной позиции в чеке клиента

customer_id	date_diff_post	startdatetime	dish_name
29891	9.0	2022-12-05 12:03:58	Кинг Фри станд
29891	9.0	2022-12-05 12:03:58	Чикен Тар-Тар
29891	9.0	2022-12-05 12:03:58	Соус Сырный
29891	9.0	2022-12-05 12:03:58	Энергет.нап. Адреналин Раш
29891	9.0	2022-12-05 14:28:35	Латте (СТАНД.)
29891	9.0	2022-12-15 00:37:19	Чизбургер
29891	9.0	2022-12-15 00:37:19	Воппер Ролл
29891	9.0	2022-12-20 09:20:38	ЧизБекон Чикен Гамбургер

Таргет одинаковый для каждого заказа

Таргет-то не настоящий!

1. В каждой строчке – по одной позиции в чеке клиента

customer_id	date_diff_post	startdatetime	dish_name
29891	0	2022-12-05 12:03:58	Кинг Фри станд
29891	0	2022-12-05 12:03:58	Чикен Тар-Тар
29891	0	2022-12-05 12:03:58	Соус Сырный
29891	0	2022-12-05 12:03:58	Энергет.нап. Адреналин Раш
29891	10	2022-12-05 14:28:35	Латте (СТАНД.)
29891	5	2022-12-15 00:37:19	Чизбургер
29891	5	2022-12-15 00:37:19	Воппер Ролл
29891	9.0	2022-12-20 09:20:38	ЧизБекон Чикен Гамбургер

Таргет одинаковый для каждого заказа

Для последней строчки — ОК

Trg = 0 -> только в последней покупке

Таргет-то не настоящий!

1. В каждой строчке – по одной позиции в чеке клиента

customer_id	date_diff_post	startdatetime	dish_name
29891	0	2022-12-05 12:03:58	Кинг Фри станд
29891	0	2022-12-05 12:03:58	Чикен Тар-Тар
29891	0	2022-12-05 12:03:58	Соус Сырный
29891	0	2022-12-05 12:03:58	Энергет.нап. Адреналин Раш
29891	10	2022-12-05 14:28:35	Латте (СТАНД.)
29891	5	2022-12-15 00:37:19	Чизбургер
29891	5	2022-12-15 00:37:19	Воппер Ролл
29891	9.0	2022-12-20 09:20:38	ЧизБекон Чикен Гамбургер

Создаем категории/подкатегории продуктов и регулярками выделяем основные признаки

Работа с исходными данными

1. В каждой строчке – по одной позиции в чеке клиента

2. Объединяем позиции в чеке

customer_id	date_diff_post	startdatetime	dish_name
29891	9.0	2022-12-05 12:03:58	Кинг Фри станд
29891	9.0	2022-12-05 12:03:58	Чикен Тар-Тар
29891	9.0	2022-12-05 12:03:58	Соус Сырный
29891	9.0	2022-12-05 12:03:58	Энергет.нап. Адреналин Раш
29891	9.0	2022-12-05 14:28:35	Латте (СТАНД.)
29891	9.0	2022-12-15 00:37:19	Чизбургер
29891	9.0	2022-12-15 00:37:19	Воппер Ролл
29891	9.0	2022-12-20 09:20:38	ЧизБекон Чикен Гамбургер

customer_id	startdatetime	dish_name
29891	2022-12-05 12:03:58	[Кинг Фри станд, Чикен Тар-Тар, Соус Сырный, Э...
29891	2022-12-05 14:28:35	[Латте (СТАНД.)]
29891	2022-12-15 00:37:19	[Чизбургер, Воппер Ролл]
29891	2022-12-20 09:20:38	[ЧизБекон Чикен Гамбургер]

Работа с исходными данными

1. В каждой строчке – по одной позиции в чеке клиента

2. Объединяем позиции в чеке

3. Идем окном по прошлым чекам и создаем фичи

Часть фичей:

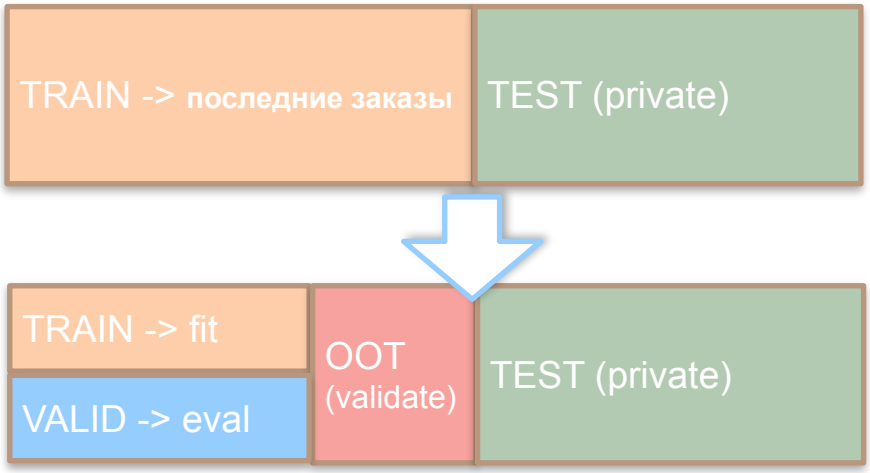
	customer_id	order_id	mean_revenue_last4	mean_revenue_last3	mean_revenue_last2	rat_revenue	mean_squares_last4	mean_squares_last3	mean_s
0	46661804	46661804_2023-08-01 18:04:56	278.3100	278.310000	287.470	1.017546	300.0	300.0	
1	52341	52341_2023-02-04 13:13:06	549.9500	549.950000	549.950	-1.000000	338.9	338.9	
2	52341	52341_2023-02-11 13:08:33	549.9500	549.950000	549.950	1.000000	338.9	338.9	

customer_id	startdatetime	dish_name
29891	2022-12-05 12:03:58	[Кинг Фри станд, Чикен Тар-Тар, Соус Сырный, Э...
29891	2022-12-05 14:28:35	[Латте (СТАНД.)]
29891	2022-12-15 00:37:19	[Чизбургер, Воппер Ролл]
29891	2022-12-20 09:20:38	[ЧизБекон Чикен Гамбургер]

1. ts: Гэпы / лаги / дни недели / время
2. items: популярные / скидки / наборы / лояльность
3. Динамика по признакам— mean/std/q25/…
4. Фичи по корзинам / число уникальных товаров + специфических (пиво, мороженное)

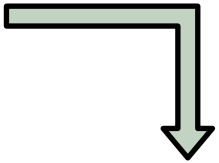
Работа с исходными данными

1. В каждой строчке – по одной позиции в чеке клиента
2. Объединяем позиции в чеке
3. Идем окном по прошлым чекам и создаем фичи
4. По каждому клиенту оставляем **последние** заказы + сплитим train/test/oot

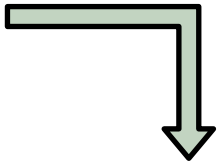


Работа с исходными данными

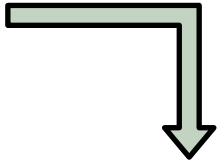
1. В каждой строчке – по одной позиции в чеке клиента



2. Объединяем позиции в чеке



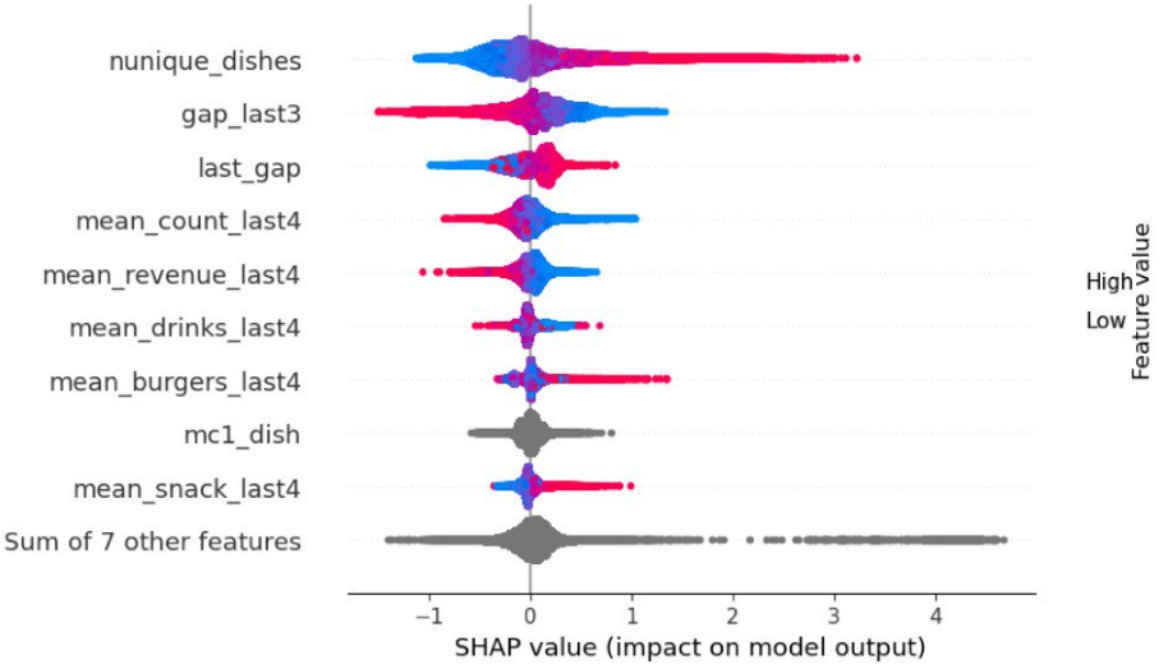
3. Идем окном по прошлым чекам и создаем фичи



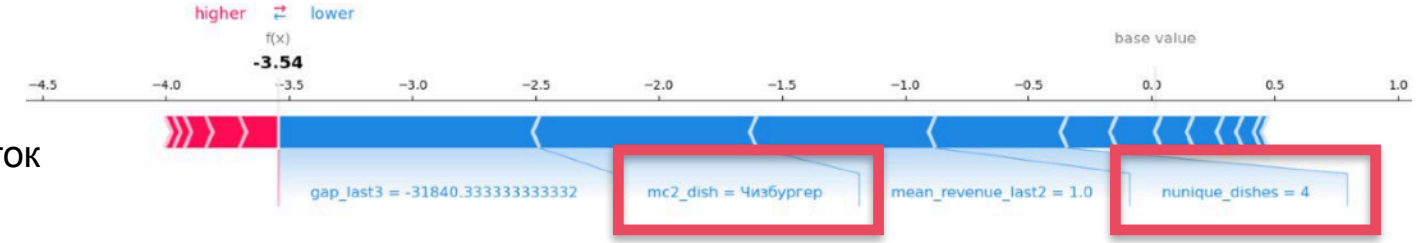
4. По каждому клиенту оставляем **последние** заказы + сплитим train/test/oot

Что влияет на отток?

- Кол-во уникальных блюд в заказе
- Длительность промежутков между заказами
- Средняя выручка с чека
- Среднее кол-во напитков/ бургеров/закусок в чеке
- Наличие в чеке самых популярных («ходовых») товаров



Конкретные примеры



Отток

Популярные товары влияющие на отток

Не отток

Число уникальных позиций

Инсайты

- ↑ уникальных позиций в чеке — ↓ отток
- ↓ дней с последнего заказа — ↓ отток
- ↓ средний чек и кол-во позиций — ↑ отток
- ↑ покупал бургеров и снеков — ↓ отток

Показатели значимости признаков (feature_importance)

	feature_importance	feature_names
0	18.263032	nunique_dishes
1	13.050467	gap_last3
2	7.852602	last_gap
3	7.147137	mean_count_last4
4	6.416983	mean_revenue_last4
5	5.842888	mean_drinks_last4
6	4.838572	mc2_dish
7	4.804129	mc1_dish
8	4.741098	rat_revenue
9	4.247841	mean_burgers_last4
10	4.079141	revenue_curr
11	4.027180	square_curr
12	3.914042	mean_snack_last4
13	3.865001	mean_revenue_last2
14	3.544248	mean_offer_last4
15	3.365639	mean_chicken_last4

CatBoostRegressor

Регрессия

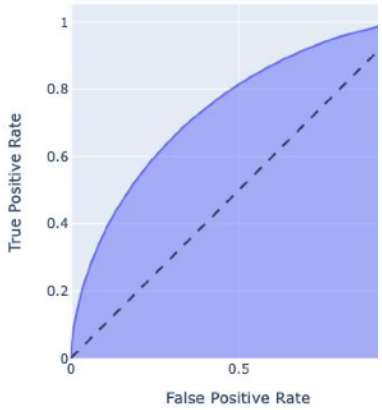
Топ признаки идентичны

Бинарная классификация

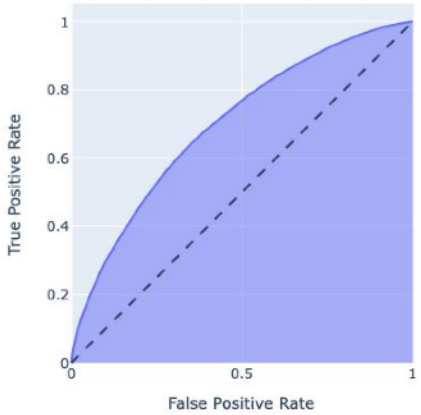
CatBoostClassifier

	feature_importance	feature_names
0	26.619333	nunique_dishes
1	14.181740	gap_last3
2	10.506269	last_gap
3	8.892528	mean_count_last4
4	6.886974	mean_revenue_last4
5	6.593787	mean_drinks_last4
6	4.621135	rat_revenue
7	4.556413	mean_burgers_last4
8	4.231381	mean_offer_last4
9	3.842078	mean_snack_last4
10	3.267994	mean_chicken_last4
11	3.028961	mc1_dish
12	2.771407	mc2_dish

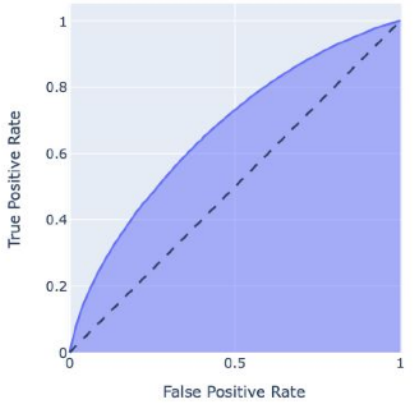
ROC Curve train (AUC=0.7420)



ROC Curve test (AUC=0.7001)



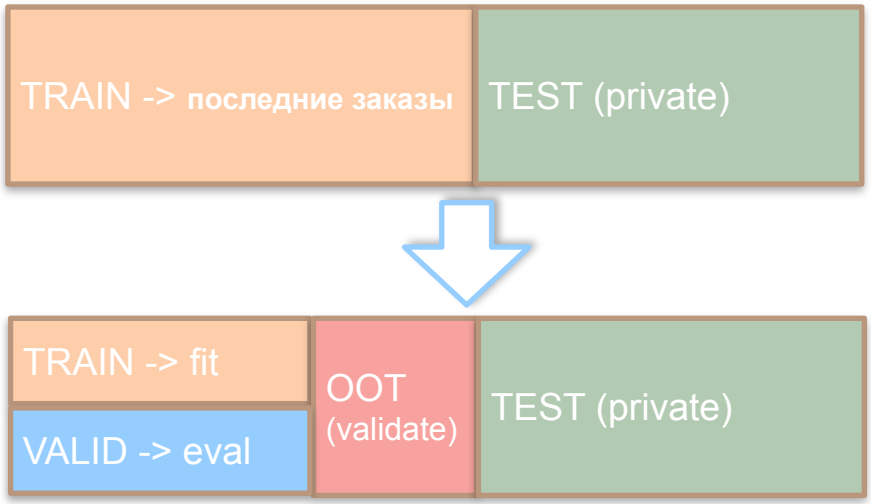
ROC Curve oot (AUC=0.6709)



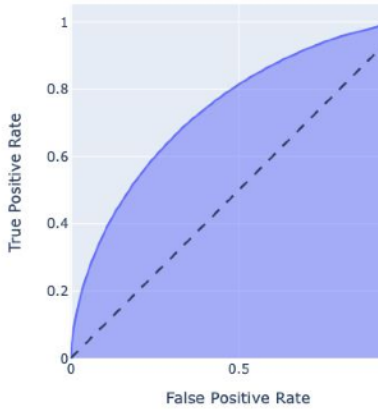
Почему скор устойчивый?

Использование только последних заказов

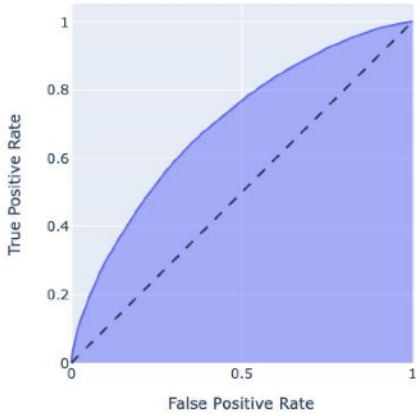
ВАЛИДАЦИЯ



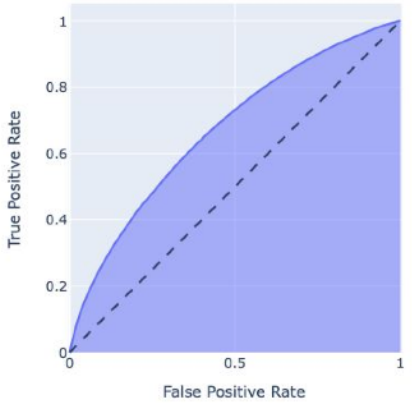
ROC Curve train (AUC=0.7420)



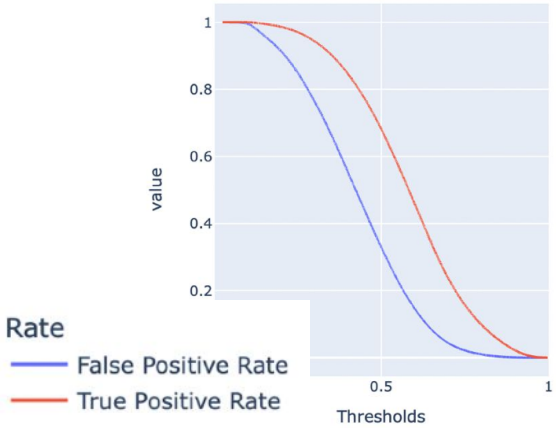
ROC Curve test (AUC=0.7001)



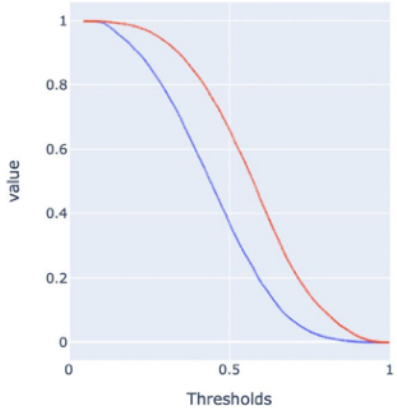
ROC Curve oot (AUC=0.6709)



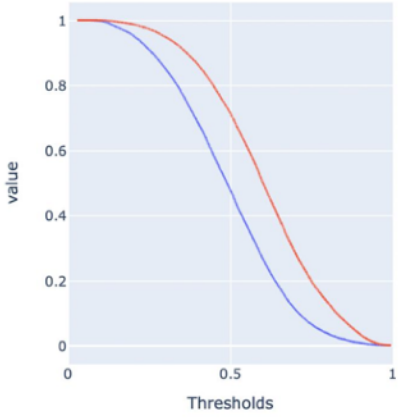
train: TPR and FPR at every threshold



test: TPR and FPR at every threshold

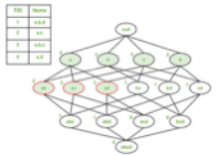


oot: TPR and FPR at every threshold



Стек технологий

TF-IDF



jupyter

python

plotly

XGBoost

scikit learn

pandas

Стоп-кодинг близко...
Что могло бы улучшить результат?

- Оптимизация фичей под число последних заказов
- Создание более детальных эмбеддингов продуктов (W2V), их совместная встречаемость (frequent itemsets)
- Выделение сезонных признаков (летом покупают мороженное)
- Выделение сегментов пользователей и обучение отдельных моделей

Демонстрация решения

Наш репозиторий на Github

Скринкаст:

Спасибо за внимание!

Ваши вопросы
и предложения?

@htklf

Наш репозиторий на Github

Скринкаст: