

Dabt

Determining the **a**ge of **b**anking
transactions

Сунгуров Владислав

Черных Фёдор

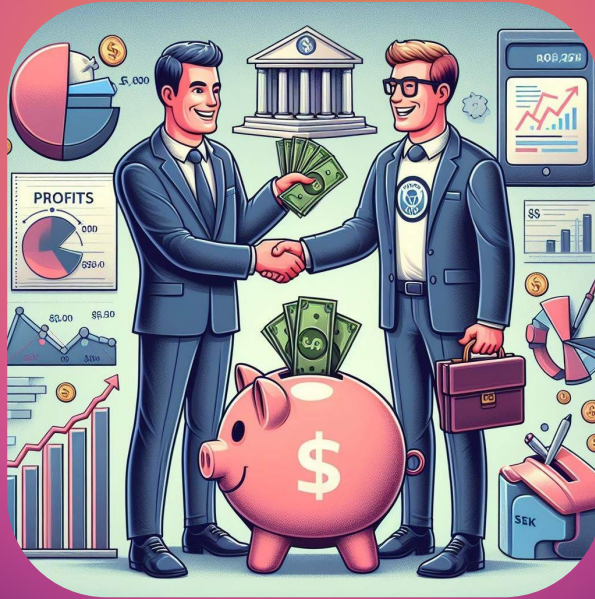
Цель проекта

Попытаться предсказать возраст пользователей по их транзакциям в интернете, чтобы помочь подобрать рекламу.

Это реальная задача с хакатона

Целевая аудитория КОДИИМ

Банки & рекламные агентства, которые с ними сотрудничают



Описание (таймлайн) работы над проектом

- Подготовка данных
- Выбор модели
- Продумывание фич
- Выбор модели
- Попытки реализации разных идей

Описание датасета

- transactions_train
- train_target
- transactions_test
- test_id

```
[ ] transactions_train.head()
```

	client_id	trans_date	small_group	amount_rur
0	33172	6	4	71.463
1	33172	6	35	45.017
2	33172	8	11	13.887
3	33172	9	11	15.983
4	33172	10	11	21.341

- client_id - уникальный идентификатор клиента
- trans_date - дата совершения транзакции
- small_group - категория покупки
- amount_rur - сумма транзакции

Описание датасета

КОДИИМ

- transactions_train
- train_target
- transactions_test
- test_id

```
train_target.head(5)
```

	client_id	bins
0	24662	2
1	1046	0
2	34089	2
3	34848	1
4	47076	3

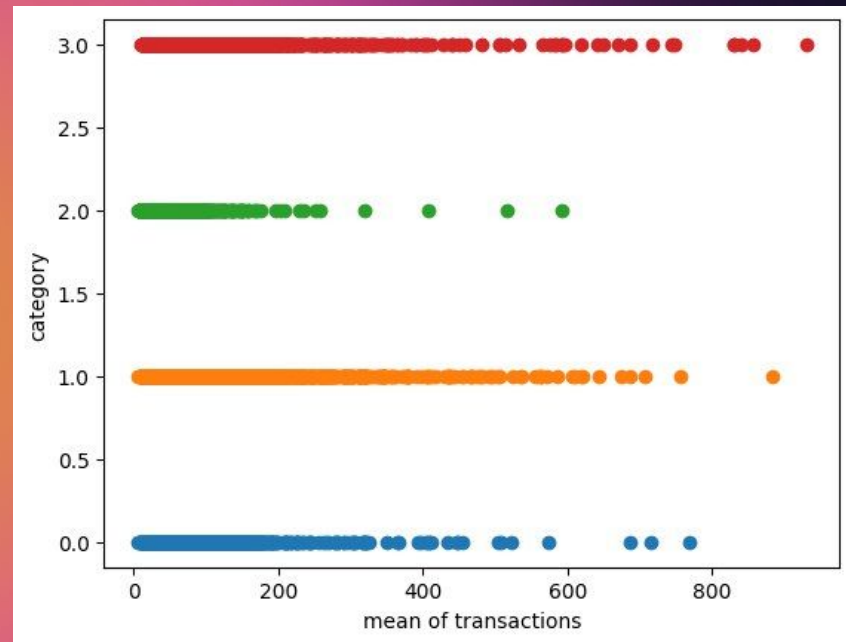
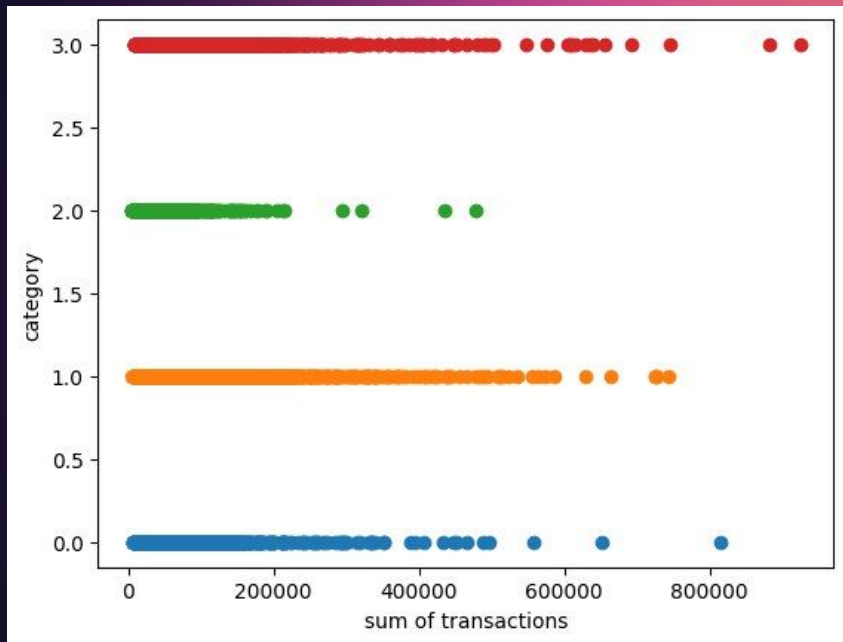
- client_id - уникальный идентификатор клиента, соответствует полю client_id из транзакций
- bins - целевая переменная, которую нужно предсказать, это категория возраста клиента

Описание датасета

- transactions_train
- train_target
- transactions_test
- test_id

То же самое что и train, только без ОТВЕТОВ

КОДИИМ



Описание препроцессинга данных

КОДИИМ

1. Считаем данные по транзакциям и правильные ответы.
2. Посчитаем по каждому клиенту самые простые агрегационные признаки. (sum, mean, std, min, max)
3. Посчитаем для каждого клиента количество транзакций

```
train.head()
```

	client_id	bins	sum	mean	std	min	max	small_group_0
0	24662	2	30254.011	34.774725	72.037354	0.074	1227.314	0.0
1	1046	0	42548.570	52.015367	106.540962	0.550	1210.506	1.0
2	34089	2	26842.816	34.325852	59.927450	0.043	782.641	0.0
3	34848	1	15773.126	16.160990	14.224936	0.043	109.590	0.0
4	47076	3	12488.375	15.929050	35.473591	0.432	541.165	0.0

5 rows × 209 columns

Модели которые мы использовали

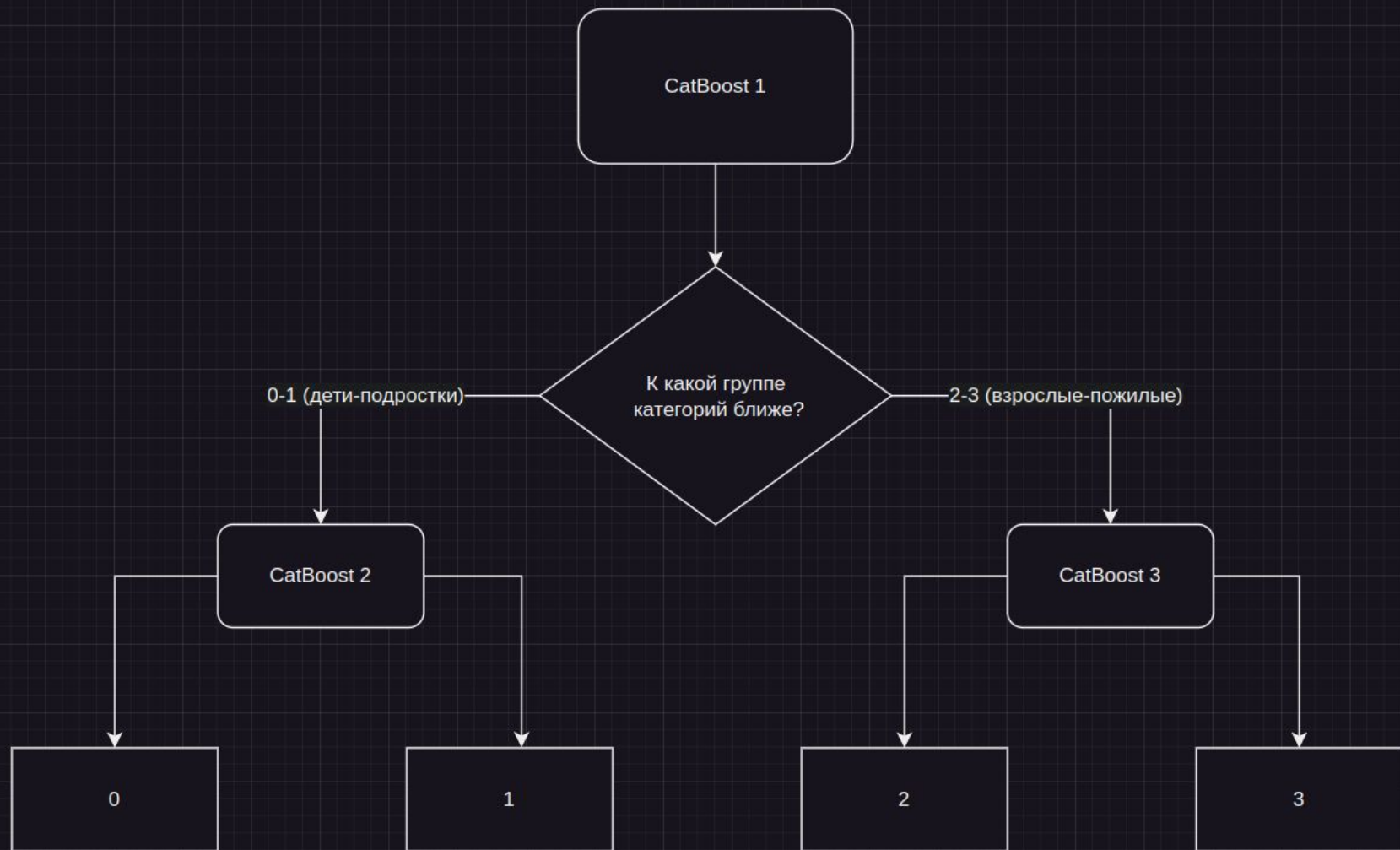
- **CatBoost 62.5~63% ⇐**
- Logistic Regression 62.5%
- Random Forest 40~50%
- Decision Tree Classifier 40~45%

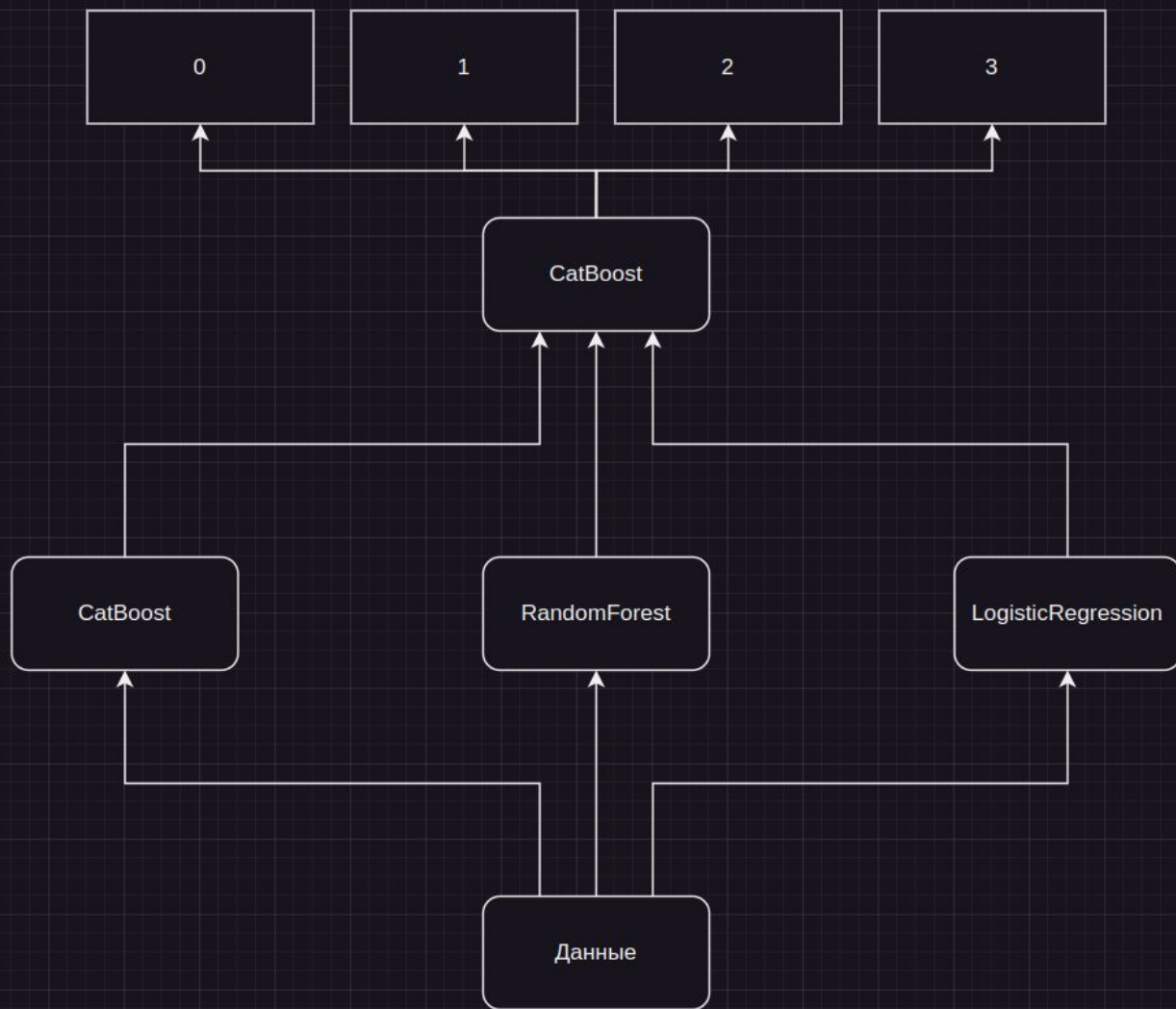
Метрика - accuracy













Перебор гиперпараметров

- Дерево: перебор глубины, минимального кол-ва листьев, условий сплита(min samples leaf, min samples split)
- Логистическая регрессия: регуляризация, итерации, сила регуляризации
- Лес: кол-во, глубина и параметры деревьев
- CatBoost: итерации, learning rate





#	Команды (24)		Финальная точность ?	Последний	Решений
1	No shakeup allowed		0,6492	4y	16
2	ГБОУ БИЮЛИ		0,6462	4y	5
3	ENOT		0,6421	4y	1
4	CPC.tomsk		0,6408	4y	8
5	Инженерная школа 1581		0,6359	4y	10
6	kore ga... requiem da		0,6352	4y	2
7	Talentum et triumphum (5)		0,6281	4y	4
8	Игорёк, прокуратура		0,6273	4y	11
9	RS		0,6197	4y	6
10	workcomeon		0,6167	4y	13



КОДИИМ