Решение задачи "Предсказание затрат"

Амир Мирас

МГУ имени М.В. Ломоносова

4 октября 2017 г.

Постановка задачи

- По статистике визитов клиентов и потраченных ими сумм необходимо предсказать категорию следующей траты для каждого клиента.
- Метрика: $\mathrm{accuracy}$
- Пусть $n = 110000$ – количество уникальных клиентов, $d = 438$ – максимальное количество дней.

Выделение обучающей выборки

$T^d =$

	1	$\cdots$	$d-7$	$d-6$	$d-5$	$d-4$	$\cdots$	d
1	9	$\cdots$	6	1	6	1	$\cdots$	0
2	0	$\cdots$	7	0	0	1	$\cdots$	0
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	0	$\cdots$	10	0	2	0	$\cdots$	0

$T^{d-7} =$

	1	$\cdots$	d
1	9	$\cdots$	6
2	0	$\cdots$	7
$\vdots$	$\vdots$		$\vdots$
n	0	$\cdots$	10

,

$Y^{d-7} =$

	Ответ
1	1
2	1
$\vdots$	$\vdots$
n	2

Обозначим данную операцию через S: $T^{d-7}, y^{d-7} = \mathrm{S}(T^d)$

План решения

1. На основе выборок T^{d-7} и T^d придумать признаки для каждой строки и получить матрицы клиентов-признаков X^{d-7} и X^d
2. Обучить модель машинного обучения на выборке $\{x_i, y_i\}_{i=1}^N, x_i \in X^{d-7}, y_i \in Y^{d-7}$
3. Предсказать ответы для выборки X^d

$$T^{d-7} = \begin{array}{c|ccc} & 1 & \cdots & d \\ \hline 1 & 9 & \cdots & 6 \\ 2 & 0 & \cdots & 7 \\ \vdots & \vdots & & \vdots \\ n & 0 & \cdots & 10 \end{array}, \quad Y^{d-7} = \begin{array}{c|c} & \textit{Ответ} \\ \hline 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ n & 2 \end{array}$$

Примитивное решение

1. Для каждого клиента посчитаем следующие признаки:
   1.1 Среднее, среднее по ненулевым, мода по ненулевым, количество нулей на всем временном ряде
   1.2 Среднее, среднее по ненулевым, мода по ненулевым, количество нулей на последней неделе
2. Обучим на этом классификатор $\mathrm{xgboost}$

На public leaderboard: **0.38206**

Валидация

- $T_{test}, Y_{test} = \mathrm{S}(T^d)$
- $T_{train}, Y_{train} = \mathrm{S}(T_{test})$
- Будем обучаться на $\{T_{train}, Y_{train}\}$ и проверять качество алгоритма на $\{T_{test}, Y_{test}\}$

Более сложные признаки

1. Доля суммы k-го класса на всем ряде T_i
2. Среднее, среднее по ненулевым, мода по ненулевым, количество нулей и доля суммы k-го класса на первых покупках каждой недели
3. Среднее, среднее по ненулевым, мода по ненулевым, количество нулей на последних двух и трех неделях
4. Вероятность первого посещения магазина для каждого дня недели
5. Последняя сумма покупки, количество дней после последней покупки и после покупки на сумму k-го класса

Обучаем на этих признаках $\mathrm{xgboost}$: **0.39775** на CV

Весовые схемы

Будем считать признаки с учетом следующей весовой схемы:

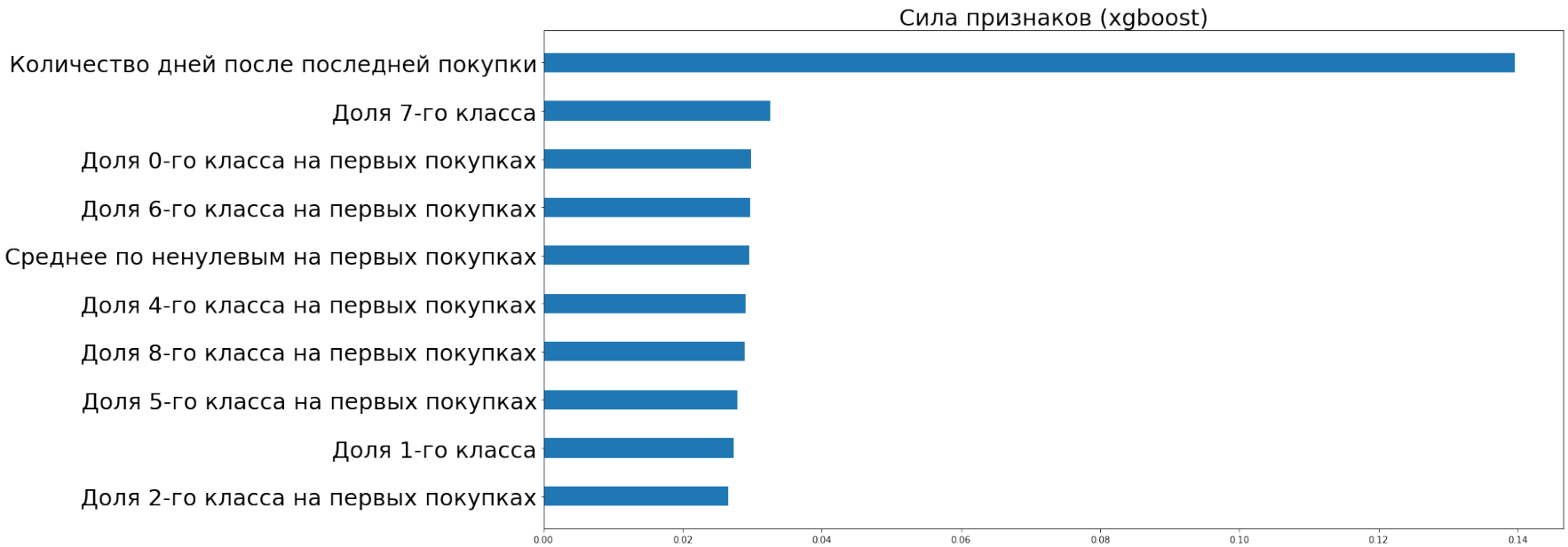
$$w_j = \left(\frac{j}{\left[\frac{d}{7}\right]} \right)^{\delta}, j = \{1, 2, 3, \cdots, \left[\frac{d}{7}\right]\} \tag{1}$$

Например, среднее:

$$\mathrm{mean}(T_i) = \sum_{j=1}^{d} w_j \sum_{m=1}^{7} T_{i,7(j-1)+m}, \tag{2}$$

то есть суммы $j-$ой недели учитываются с весом w_j.
Обучаем $\mathrm{xgboost}$: **0.3987** на CV

Топ 10 признаков



Выбор алгоритма

Алгоритм	**Качество на CV**	**Время работы**
xgboost	0.3987	6.45 мин
lightgbm	**0.3989**	**1.76 мин**

Ансамбль

Рекурсивным путем посчитаем выборки

$$T^{d-7}, Y^{d-7} = \mathrm{S}(T^d), T^{d-14}, Y^{d-14} = \mathrm{S}(T^{d-7}) \quad (3)$$

$$T^{d-21}, Y^{d-21} = \mathrm{S}(T^{d-14}), T^{d-28}, Y^{d-28} = \mathrm{S}(T^{d-21}) \quad (4)$$

- Обучим $\mathrm{lightgbm}$ на $\{T^d, Y^d\}$, $\{T^{d-7}, Y^{d-7}\}$, $\{T^{d-14}, Y^{d-14}\}$, $\{T^{d-21}, Y^{d-21}\}$, $\{T^{d-28}, Y^{d-28}\}$ и получим предсказания $p^d_{lgb}, p^{d-7}_{lgb}, p^{d-14}_{lgb}, p^{d-21}_{lgb}, p^{d-28}_{lgb}$
- Обучим $\mathrm{xgboost}$ на объединении $\{T^{d-7}, Y^{d-7}\}$, $\{T^{d-14}, Y^{d-14}\}$, $\{T^{d-21}, Y^{d-21}\}$, $\{T^{d-28}, Y^{d-28}\}$ и получим предсказание p_{xgb}
- Итоговая модель: $0.5p^d_{lgb} + 0.3p^{d-7}_{lgb} + 0.3p^{d-14}_{lgb} + 0.05p^{d-21}_{lgb} + 0.05p^{d-28}_{lgb} + 0.1p_{xgb}$

На CV: **0.403**
На public leaderboard: **0.40527**

Что за данные?