Предсказание популярности банкомата

Happy Data Year

22 ноября — 10 января

18.01.2019

ОГЛАВЛЕНИЕ

Обо мне

МосТрансПроект

Data Scientist, Data Science TeamLead

fasten

Data Scientist

DO

Data Scientist

Задача

Пример исходных данных:

```
In [65]: train_show.sample(12)
```

Out[65]:

	id	atm_group	address	address_rus	lat	long	target
498	8413.0	5478.0	PL. KOMENDANTSKAYA SANKT-PETERB	NaN	NaN	NaN	-0.037399
838	7014.0	5478.0	D. 1, UL. LENINA UGLICH G	улица Ленина, 1, Углич, Ярославская область, Р...	57.526166	38.315474	-0.071104
1661	2070.0	3185.5	LENINA, 40 TOMSK	проспект Ленина, 40, Томск, Россия, 634034	56.474112	84.949697	-0.006256
5091	8555.0	496.5	BORISA BOGATKOVA 201 NOVOSIBIRSK	улица Бориса Богаткова, 201, Новосибирск, Росс...	55.035364	82.973745	-0.004055
1261	1980.0	496.5	LENINA 164 LABINSK	улица Ленина, 166, Лабинск, Краснодарский край...	44.635516	40.726803	-0.029180
2573	8553.0	496.5	POBEDY 28A YAROSLAVL	улица Победы, 28А, Ярославль, Россия, 150040	57.634250	39.870763	-0.104455
1731	2558.0	3185.5	MICHURINSKIJ PR.,D7 MOSKVA	NaN	NaN	NaN	-0.049696
1155	3759.0	8083.0	PR.T KIROVA, 257 Samara	проспект Кирова, 257, Самара, Россия, 443091	53.239170	50.238462	0.191143
568	5025.0	5478.0	D. 23D, UL. MALINOVSKOGO ROSTOV-NA-DO	улица Малиновского, 23, Ростов-на-Дону, Россия...	47.226686	39.612686	-0.004407
4317	4923.0	5478.0	D. 2A, UL. KERAMICHESKAYA ULAN-UDE G	Керамическая улица, 2А, микрорайон Стеклозавод...	51.863383	107.542012	0.007746
2995	2815.0	3185.5	ST.M. ZOLOTAJA NIVA NOVOSIBIRSK	метро Золотая Нива, Дзержинская линия, Новосиб...	55.037969	82.976215	-0.018231
4485	6226.0	5478.0	D. 25-A, UL. LENINA KRASNOGORSK	улица Ленина, 25А, Красногорск, Московская обл...	55.827130	37.312010	-0.015042

Задача:

Предсказание индекса популярности геолокации для размещения устройства банкоматной сети.

В обучающей выборке находятся данные о геопозиции шести тысяч банкоматов Росбанка и его партнеров, а также целевая переменная — индекс популярности банкомата. В тестовой выборке еще две с половиной тысячи банкоматов, разделенных поровну на публичную и приватную часть.

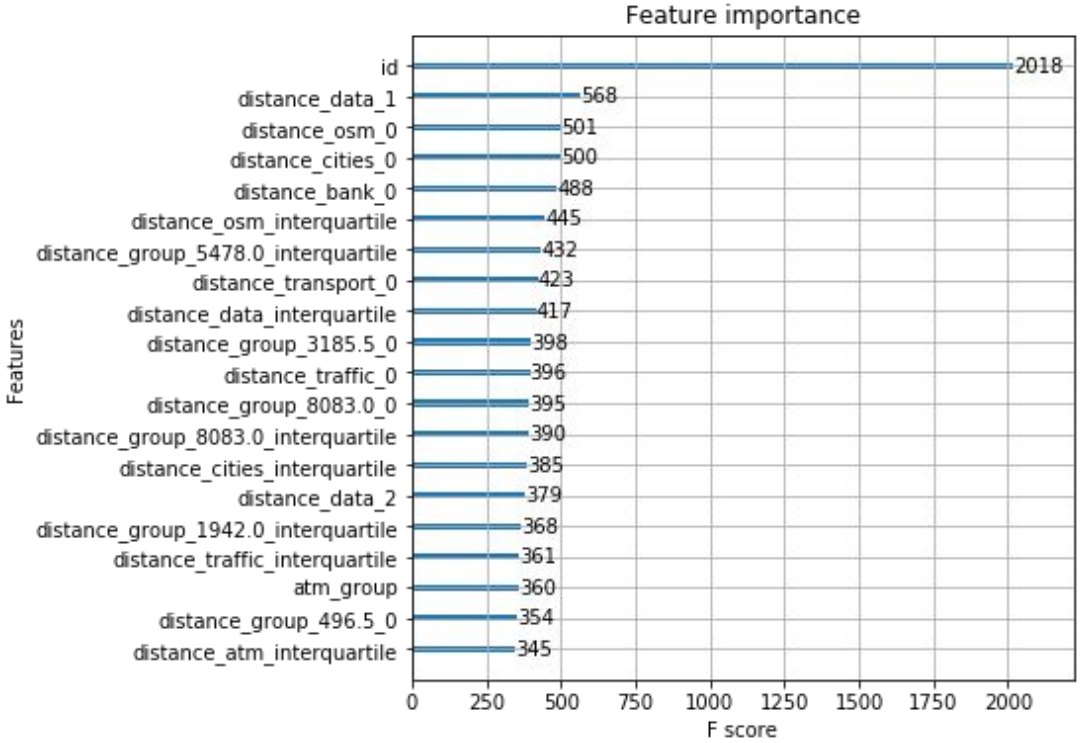
Особенности данных:

1. **Сравнительно небольшой набор данных**
2. **Наличие выбросов и “дублей”**
3. **Пропуски и неверно определенные адреса**
4. **Важность ID, как признака**
5. **Standart Scaling по train и test вместе (предположительно)**

Подходы к решению

1. **Данные:**
   a. Open Street Maps - банкоматы, отделения банков, POI (points of interest), населенные пункты, общественный транспорт
   b. Исходные данные - расстояния до ближайших банкоматов
   c. Исходные данные - расстояния до банкоматов из группы
   d. Пропуски в данных (около 300 адресов) исправлялись вручную
2. **Алгоритм:**
   a. XGboost
   b. Гиперпараметры подбирались посредством hyperopt
3. **Валидация:**
   a. 10-KFold валидация без стратификации
4. **Дополнительно:**
   a. Предсказания всех 10 фолдов усреднялись в финальном решении

Ключевые признаки и результаты



1. **Результат на валидации:**
   a. mean rmse: **0.04217**
   b. std rmse: **0.0007**
2. **Результат на публичной части:**
   a. rmse: **0.042692**
   b. место: **14**
3. **Результат на приватной части:**
   a. rmse: **0.042348**
   b. место: **15**

Что не сработало

1. **Данные:**
   a. Open Street Maps - разбиение POI на подгруппы (спорт, покупки, еда и т.п.)
   b. Фильтрация минимальных и максимальных значений
   c. Признак принадлежности к кластеру, полученному посредством KMeans
2. **Алгоритм:**
   a. Catboost (и категориальные признаки)
   b. KNN, ridge regression (взвешенная сумма с бустингами)
3. **Валидация:**
   a. 10-KFold валидация со стратификацией
4. **Дополнительно:**
   a. Классификация максимальных значений с последующим увеличением значений популярности в зависимости от вероятности быть максимумом

Дальнейшее развитие

1. **Фильтрация данных:**
   a. Использование данных только с накопленной историей (исключение “свежих” банкоматов)
   b. Исключение данных (либо добавление признака) для банкоматов с независимым от местоположения денежным потоком
   c. Исключение особых случаев (напр. В/Ч)
2. **Данные о клиентах:**
   a. Данные о наличии организаций с зарплатными клиентами в окрестности банкомата
   b. Данные о клиентах (домашний и рабочий адреса)
3. **Данные о покупках и окружении:**
   a. Данные о покупках от мерчантов в окрестности банкоматов
   b. Данные об оценке корпоративных клиентов (напр. организаций-заемщиков) в окрестности банкоматов

Happy Data Year

22 ноября — 10 января

email: arty.erokhin@gmail.com

tg: @Gofat