Александр Дроботов

sashadrbtv@gmail.com

Результаты

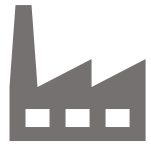
Модель	XGBoost binary logistic & regression (отдельно 4 модели)
Binary logistic (I)	Вероятность попадания транзакции в радиус 0.02 от дома/работы
Regression (II)	Дистанция в метрах от транзакции до дома/работы
Финальный результат	Объединение результатов по 4-ем моделям; отбор ближайших точек; финальный расчет широты и долготы
+DBSCAN clustering	Создания дополнительных признаков на основе кластеров для каждого ID
Процессор	Intel core i5-6500 CPU @ 3.2 GHz, x64
ОЗУ	8 ГБ
Видеокарта	Intel HD Graphics 530
Оценочное время вычисления	< 10 минут предобработка данных и расчет финального результата ≈ 1 час для расчета 4-ех моделей

Основные этапы

1 Предобработка

2 Построение моделей

3 Вычисление координат дома и работы



1. Предобработка

ID [home/work duplicates]	нек-рые ID имели > 1 дома/работы; для таких ID оставляем координаты дома и работы, которые имеют самую высокую частоту
reverse geocoding	исходные столбцы country и atm & pos address не имели единого формата; использовалась offline библиотека для обратного геокодирвоания (из координат в страну, место, субъект)*
country	сокращение кол-ва транзакций: оставляем только РФ и Украину
place & region	сокращение кол-ва транзакций: убираем места с частотой <=0.07 и регионы с частотой <=0.2 для каждого ID
ID [wo transactions near home/work]	для каждой модели рассчитывался уникальный целевой вектор; ID, у которых не было транзакций в радиусе 0.02 с домом/работой, выбрасывались из трейна

*лицензия для открытого использования

2. Построение моделей

amount	преобразование: exp(amount), 10 ** amount; классический набор статистик* по: ID, дню недели, mcc кодам, месту, региону + комбинации с ID
transaction_date	извлекаем: месяц, неделя, день, день недели, #недели в месяце, праздники и нерабочие дни; кол-во транзакций в день, кол-во дней с момента последней транзакции, кол-во дней присутствия в регионе
mcc & place & region	частота для каждого ID; отмечаем водителей
target variable	сглаживание Лапласа (additive soothing) с целевой переменной (метка 1/0 или дистанция)
DBSCAN clustering	разбиение транзакций на кластеры по координатам для каждого ID (epsilon = 2 км, min_samples = 10%); для каждого кластера находим центр и отклонение от него, считаем статистики

*статистики: sum, min, max, mean, median, count, std, variance

3. Вычисление координат дома и работы

1) Объединение результатов по моделям
2) Применение формулы: $x = probability \times \frac{1}{distance^{2}}$
3) Отбор топ *k* транзакций, в радиусе *r* от *max(x)*
4) Финальный результат – взвешенное на значение *x* среднее широты и долготы

Как применять результаты?

- автокредит, клиенту, которому далеко добираться до работы
- ипотеку, человеку, который находится в районе с максимальной плотностью клиентов с ипотекой
- страхование, клиенту, который часто выезжает за рубеж в определенное время

Как применять результаты?

- делать уникальные предложения клиентам других банков
- которые сняли деньги в нашем АТМ или оплатили покупку через наш POS-терминал
- модель дает возможность понять, где живут и работают такие люди, где максимальное сосредоточение таких людей
- в таких местах можно размещать дополнительную рекламу (к примеру, формата ООН), нацеленную на клиентов других банков