Raiffeisen Data Cup

Сергей Гайдаенко

Входные данные

- Лог транзакций (~1 200 000 штук)
  - `customer_id // 10 000`
  - `terminal_id`
  - `city`
  - `country`
  - `currency`
  - `mcc`
  - `transaction_date`
  - `amount`
  - `atm_address`
  - `atm_address_lat`
  - `atm_address_lon`
  - `pos_address`
  - `pos_adress_lat`
  - `Pos_adress_lon`
- Тренировочный набор: +`home_lat, home_lon, work_lat, work_lon`

Задача

- Определить домашний и рабочий адрес клиента с точностью 0.02 градуса
  - ~1-2 км
- Оценка - доля успешно определенных адресов

Схема

- Объект классификации - точка с координатами
  - Транзакция
  - { customer_id; terminal_id }
  - Произвольная точка
- Две модели: для дома и работы
  - Признаки одинаковые
- Target
  - 0.0 для точки вне целевой окрестности
  - (0.0, 1.0] для точки в целевой окрестности (линейная зависимость)
- Catboost
  - Метрики регрессии: RMSE, MAE, CrossEntropy
- Адрес - координаты точки с лучшей оценкой

Фильтрация

- terminal_id - не подошел
  - кластеризовал по MCC с радиусом ~0.003
- Код страны - только "RU"/"RUS"
- Код валюты - только 643.0
- city, address - грязные
- amount имеет разный смысл для POS и ATM
- Точки не дальше 0.4 градуса от центра
  - В идеале - сетка

Фильтрация

```
customer_id
terminal_id
city
country
currency
mcc
transaction_date
amount
atm_address
atm_address_lat
atm_address_lon
pos_address
pos_adress_lat
pos_adress_lon
```

```
customer_id
merchant_id
mcc
transaction_date
amount_pos
amount_atm
transaction_lat
transaction_lon
```

Признаки

- Расчет по окрестности точки

- Чтобы наверняка, по 4 окрестностям:
  - R = 0.1
  - R = 0.05
  - R = 0.02
  - R = 0.01

Признаки: количество транзакций

```
// Количество транзакций клиента в окрестности.
FF_CustomerTransactionCount,

// Доля транзакций клиента в окрестности среди всех
транзакций клиента.
FF_CustomerTransactionRate,

// Доля транзакций клиента в окрестности среди всех
транзакций торговых точек в окрестности.
FF_MerchantTransactionRate,
```

Признаки: суммы транзакций (1/2)

```
// Доля средств, потраченных клиентом в окрестности, среди
всех потраченных клиентом средств.
FF_CustomerAmountRate,

// Доля наличных, снятых в банкоматах в окрестности, среди
всех снятых клиентом в банкоматах наличных.
FF_CustomerAmountAtmRate,

// Доля средств, потраченных клиентом в окрестности, среди
всех потраченных средств в окрестности.
FF_MerchantAmountRate,
```

Признаки: суммы транзакций (2/2)

```
// Средняя сумма транзакции в окрестности.
FF_AvgAmount,

// Среднеквадратичное отклонение суммы транзакции.
FF_VarAmount,

// То же самое, но для снятия наличных в банкоматах.
FF_AvgAmountAtm,
FF_VarAmountAtm,
```

Признаки: объекты (1/2)

```
// Количество торговых точек в окрестности.
FF_MerchantCount,

// Количество торговых точек в окрестности, где были
транзакции клиента.
FF_MerchantUsedCount,

// И их доля по отношению к общему количеству.
FF_MerchantUsedRate,
```

Признаки: объекты (2/2)

```
// Количество известных домашних адресов в окрестности.
FF_HomeCount,

// Количество известных рабочих адресов в окрестности.
FF_WorkCount,

// Отношение количества известных домашних/рабочих адресов к приблизительному количеству активных клиентов в окрестности.
FF_HomesPerCustomer,
FF_WorksPerCustomer,
```

(Эх, надо было и просто количество активных клиентов добавить...)

Признаки: время

```
// Временной охват транзакций в окрестности (макс -  мин)
FF_TransactionPeriod,

// По окрестности / по всей истории
FF_TransactionPeriodRate,

// Количество месяцев/недель/дней в году, когда были
транзакции.
FF_MonthCount, // [1..12]
FF_WeekCount, // [1..365/7]
FF_DayCount, // [1..365]

// По окрестности / по всей истории.
FF_DayCountRate,
```

Признаки: выходные

```
// Доля транзакций клиента в окрестности, совершенных в выходные.
FF_DayOffTransactionRate,

// Доля наличных средств, снятых в окрестности в выходные.
FF_DayOffAmountAtmRate,
```

С учетом официальных дат праздников в 2017 году.

Признаки: MCC

```
// Суммарный вес торговых точек в окрестности,
// где были транзакции клиента, с учетом вероятности,
// что торговая точка недалеко от домашнего/рабочего адреса.
//
// P_mcc_rate_in_target_area / P_mcc_rate_global - 1.0
//
FF_HomeUsedMccWeight,
FF_WorkUsedMccWeight,
```

Признаки: не окрестные

```
// Расстояние от центра транзакций клиента до точки.
FF_DistanceFromCenter,

// Широта/долгота.
FF_Lat,
FF_Lon,
```

Признаки: top

```
// Самые полезные - в версиях для окрестности 0.01 и 0.02
FF_CustomerAmountRate,
FF_CustomerAmountAtmRate,
FF_CustomerTransactionRate,
FF_DayCountRate,
FF_DayOffTransactionRate,
FF_HomeUsedMccWeight,
FF_HomesPerCustomer,
FF_WorksPerCustomer,
FF_MerchantUsedCount,
FF_MerchantUsedRate,
FF_DistanceFromCenter,
```

Признаки: резюме

- Итого ~130 признаков
- Большинство, скорее всего, бесполезные
- Можно придумать больше и лучше
- Brainstorming

Валидация

- ...
- Переобучение
- Отбор признаков

Инструменты

- Compaq Presario CQ61
  - AMD Turion II M500 2200 Mhz
  - 4 GB RAM
- catboost-0.6.3.exe
  - и немного cmd-скриптов
- Тестовая модель: 100 итераций, ~15 минут
  - запись признаков в csv через iostream - *FUUUU.png*
- "Релизная" модель: 4000 итераций, ~6 часов

Потраченное время

- Много

Вопросы?

- Github - coming soon
  - Надо привести код в порядок
  - Метод ExtractFeatures - ~500 строк