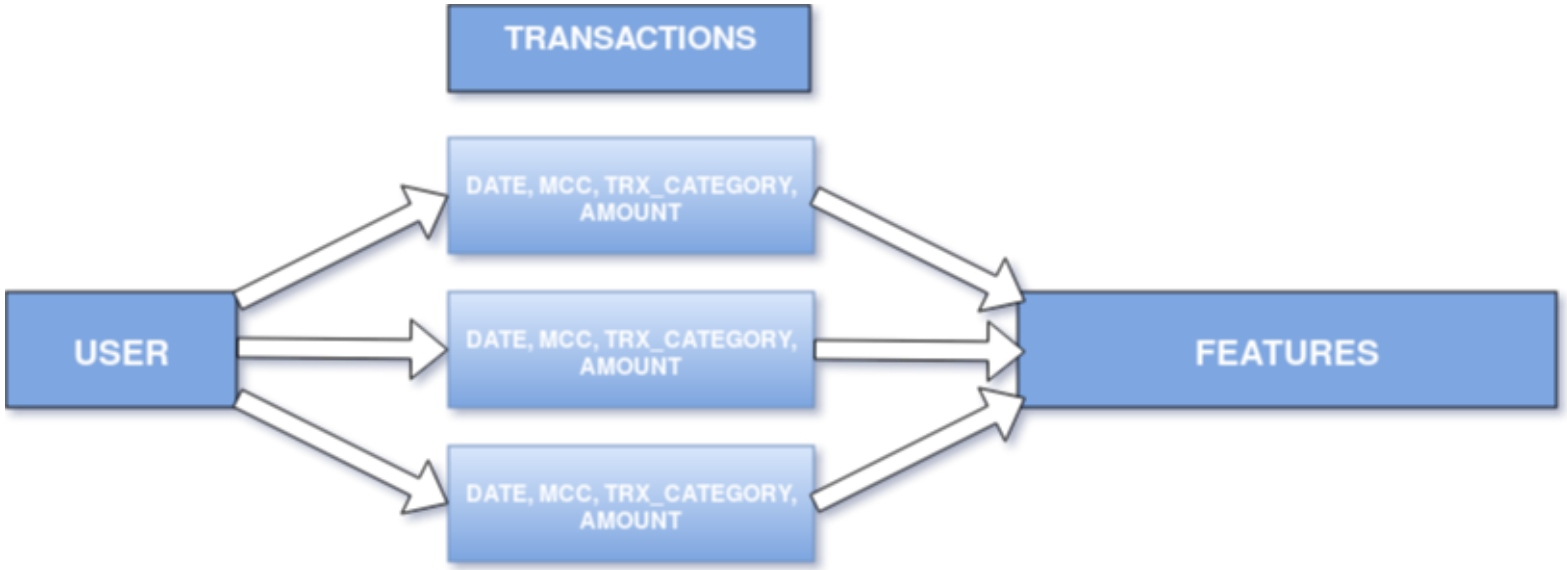
Rosbank ML Competition

Хрыльченко Кирилл

1 июня 2018 г.

Постановка задачи

- **Данные:** истории банковских транзакций по «Сверхкарте+» за льготный период.
- **Задача 1:** оценка вероятности продолжения пользования картой.
- **Задача 2:** прогноз суммарных трат в POS-терминалах за следующие три месяца пользования.
- **Структура данных:**



..

Временные признаки

- Временные — **лучшие** признаки.
- **ROC-AUC**: 0.851 – 0.856 (Stratified Kfold на train'e)
- Промежутки времени приводятся к части года: $\frac{T}{365}$
- Обработка группы дат:
  - Начало, конец, продолжительность пользования.
  - *mean, median, max, min, std* для промежутков времени между последовательно идущими датами.
- Деление на группы:
  - По времени:
    - Всё время пользования.
    - Первый календарный месяц.
    - Последний календарный месяц.
    - Последние две недели.
  - По типу транзакций (11 групп — всё вместе + каждый тип отдельно).

Денежные признаки

- **ROC-AUC**: 0.831 – 0.841
- Обработка группы трат: *mean*, *median*, *max*, *min*, *std*, *sum*.
- Деление на группы:
  - По времени:
    - Всё время пользования.
    - Последний календарный месяц.
  - По типу транзакций:
    - Все транзакции.
    - Транзакции, взятые со знаком.
    - Каждый тип транзакции отдельно.

Остальные признаки

- Валюты, типы транзакции и MCC-категории:
  - Количество уникальных вхождений.
  - Доли всех различных возможных типов признака в истории.
- Канал привлечения клиента — категориальный признак.
- ID клиента.

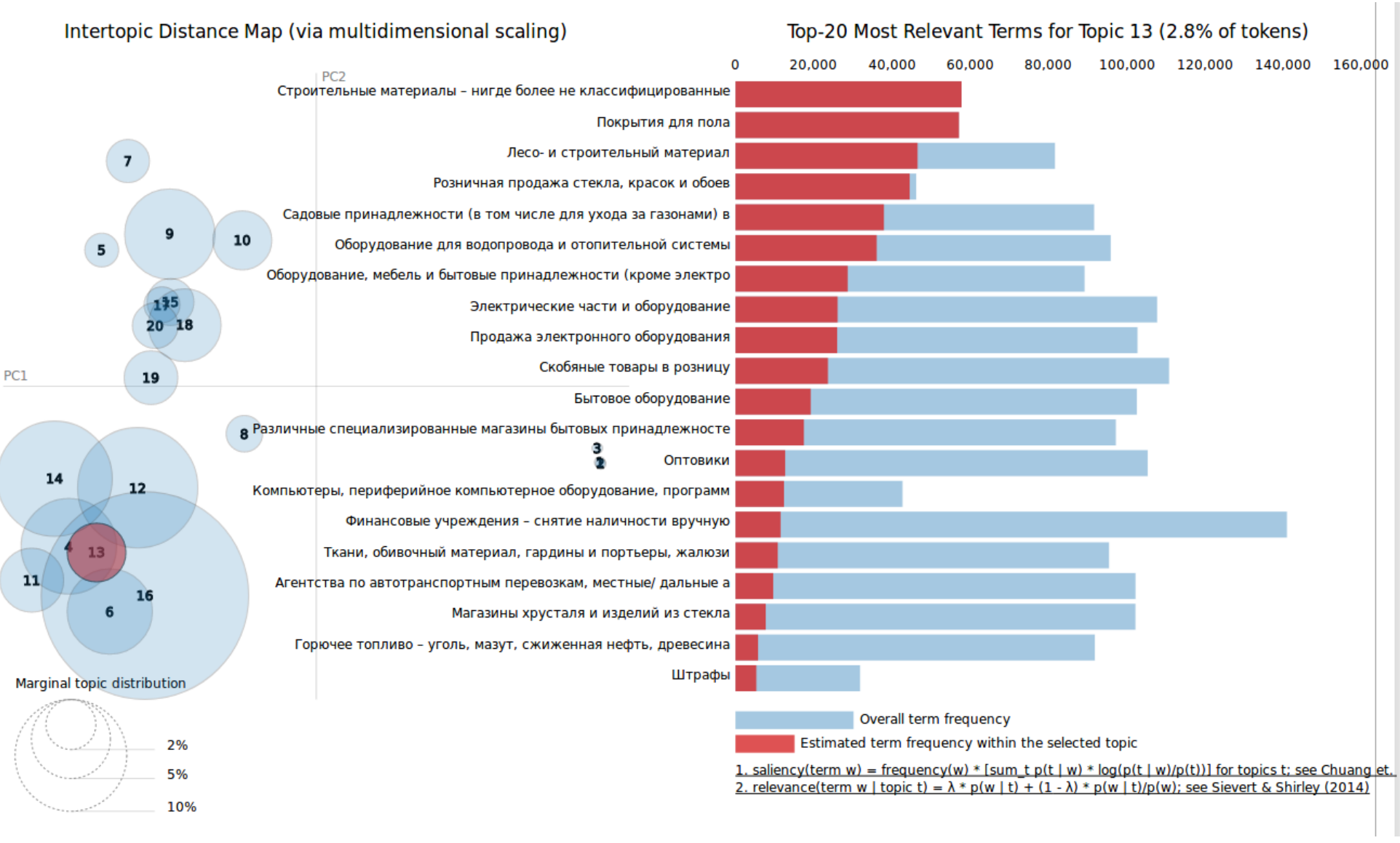
Модели

- **Классификация:**
  - Одна модель *lightgbm*, обучаемая на всех тренировочных данных.
  - subsample = 0.5, colsample bytree = 0.8, learning rate = 0.01.
  - 800 деревьев — среднее количество деревьев до early stopping на CV + 100.
- **Регрессия:**
  - Усреднение предсказаний моделей *lightgbm*, обученных на CV.
  - Параметры такие же, как у классификатора.
  - early stopping rounds = 200.
- **Private scores:**
  - task 1, best single model: 0.8743 (1st place)
  - task 1, blend with tearth's baseline: 0.8747 (1st place)
  - task 2, best single model: 3.8555 (2nd place)

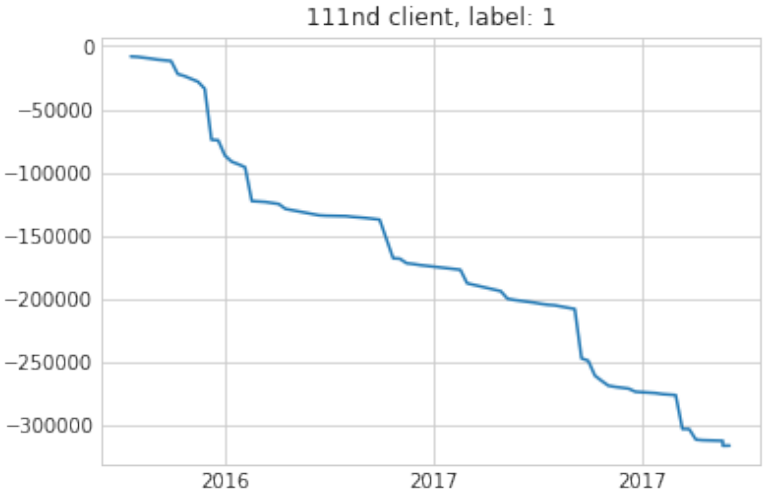
Анализ MCC-кодов

- Деление на группы — это хорошо:
  Временные признаки без деления по типу транзакций и с делением: $0.826 \rightarrow 0.851$
- $\Rightarrow$ Надо делить на группы по MCC-кодам.
- Способы деления на группы:
  - Сайт mcc-codes.ru — «плохое» деление.
  - Тематическое моделирование.
    - **ROC AUC** временных признаков: 0.8556
    - **ROC AUC** денежных признаков: 0.8416
    - Выявляет такие группы, как:
      туризм, медицина, одежда/семья, строительство/ремонт, финансы, автомобили/транспорт, спорт.
    - Анализ векторных представлений MCC-кодов с помощью рекуррентных нейронных сетей.

Тематическое моделирование



Бонус



- Анализ баланса клиента:
  - Автокорреляция.
  - Сглаженность - $\frac{mean}{std}$.
  - Монотонность.

- Для анализа массивов можно и нужно использовать квантили (например, с помощью *np.percentile*).
- Анализ логарифмических приращений трат.
- Скрытая марковская модель годится для анализа MCC-кодов, трат и почти чего угодно!

Спасибо за внимание!