Войти в топ 34, ничего не делая. Инструкция по применению.

Викулин Всеволод

МГУ имени М.В. Ломоносова

va.vikulin@physics.msu.ru

24 ноября 2016 г.

Самый важный слайд

Шаблон доклада в одной картинке:

Осталось разобраться с ???

Три задачи на «Сбербанке»

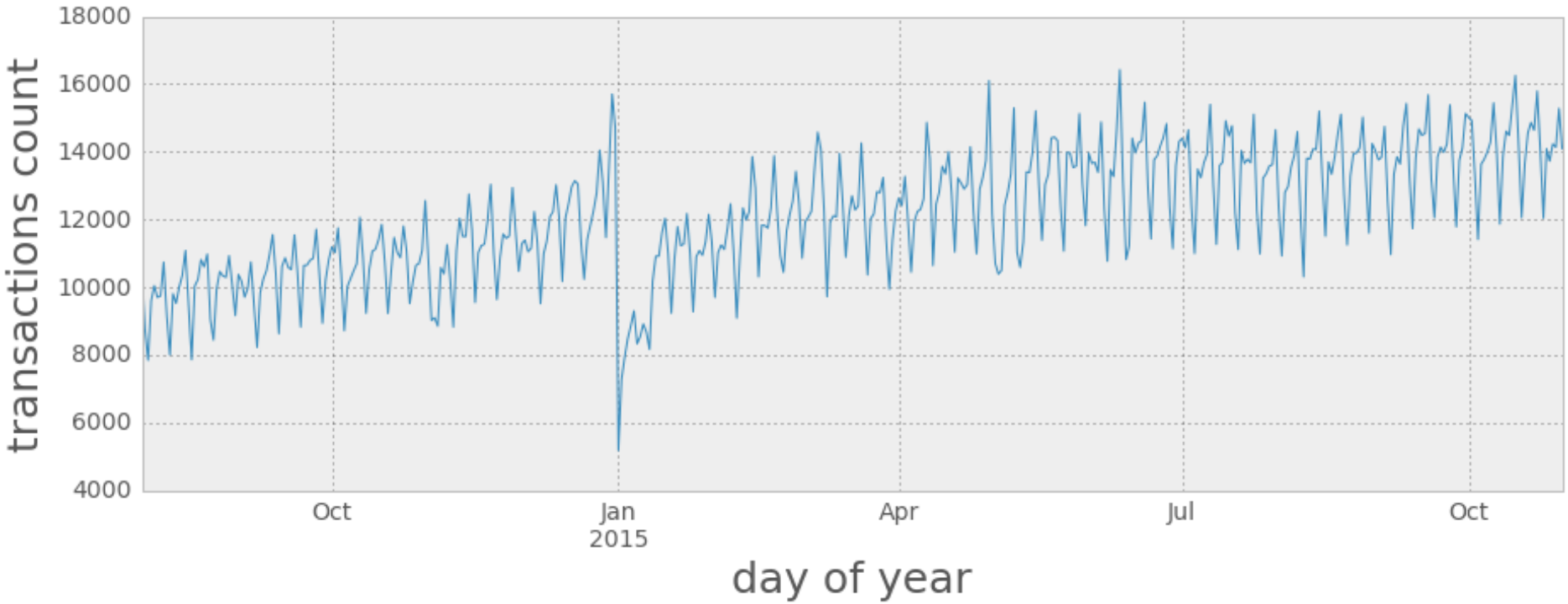
Богатый выбор из трех задач:

- Предсказать вероятность мужского пола (200 баллов)
- Предсказать объем трат по каждой из 184 категорий на каждый день следующего месяца (300 баллов)
- Предсказать объем трат в следующем месяце в каждой из 184 категорий для каждого customer (400 баллов)

Шаг первый

Решаем нормально только самую дорогую задачу, на остальные не тратим время.

Количество покупок по дням



Шаг второй

Немножко (совсем немножко) смотрим на данные

Шифрованные суммы

Топ 10 по встречаемости положительных транзакций

1) 22459 2) 112295 3) 44918 4) 67377 5) 11229 6) 224591 7) 2245 8) 89836 9) 33688 10) 4491

Преобразованные суммы

Топ 10 по встречаемости положительных транзакций

1) 1000 2) 5000 3) 2000 4) 3000 5) 500 6) 10000 7) 100 8) 4000 9) 1500 10) 200

Метрика

Регрессия. Метрика - *RMSLE*.

Метрика

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(\log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

Шаг третий

Настраиваемся под метрику

Валидация

В большинстве задач приходится думать над валидацией.
Просто **from sklearn.model-selection import train-test split** не лучший вариант.
Здесь лучше всего работала валидация на n последних известных месяцах.

Шаг четвертый

Нормально валидируемся.

Признаки

Группируем данные по (customer, mcc, месяц). Признаки:

- Прошлые суммы покупок пользователя (прошлые месяцы)
- Прошлогодние суммы покупок
- Средние (медианы) суммы покупок по данному пользователю, данному mcc
- Средняя (медиана) суммы покупок по данному типу транзакции и данному терминалу
- Дополнительно выкачиваем погоду и курс рубля.
- Еще много чего по мелочи, но жизнь слишком коротка, чтобы это слушать

Шаг пятый

Вытаскиваем нормальные признаки

Почему я люблю XGBoost

- Универсальный аппроксиматор из коробки
- Нормально написан (быстрый, многопоточный)
- Легко поставить, интерфейс как в Scikit learn
- Много параметров, есть что улучшать.

Заключение

Общее время работы два дня, около 12 человек-часов. 34 место по всему лидерборду, 16 место по 3 задаче. Бустите ваши алгоритмы (пока Евгений не видит)