SNAHackathon

Мурашкин Вячеслав

Задача

- Предоставлен граф друзей для миллиона пользователей
- Часть связей внутри миллиона удалена
- Задача – найти удаленные связи

Метрика

- Метрика: nDCG усредненный по пользователям
- Ограничение: в сабмите не более 4M связей

Метаданные

- Типы связей: коллеги, сослуживцы и тд
- Демография: пол/возраст
- География: страна, регион из профиля

Baseline

- Логистическая регрессия предсказывающая вероятность связи на основе числа **общих друзей** и совпадения пола
- В обучение попадают пары имеющие общих друзей

Baseline

- Не подбирался коэффициент для intercept – критично для задач в которых классы **не сбалансированы**
- Чтобы пройти baseline достаточно убрать **фильтры** на число общих друзей у кандидатов – **повысить полноту**

Что пробовали

- Соцдем статистика по друзьям
- Поиск сообществ – BigCLAM
- SVD разложения для географии
- ALS для получения латентных факторов пользователей

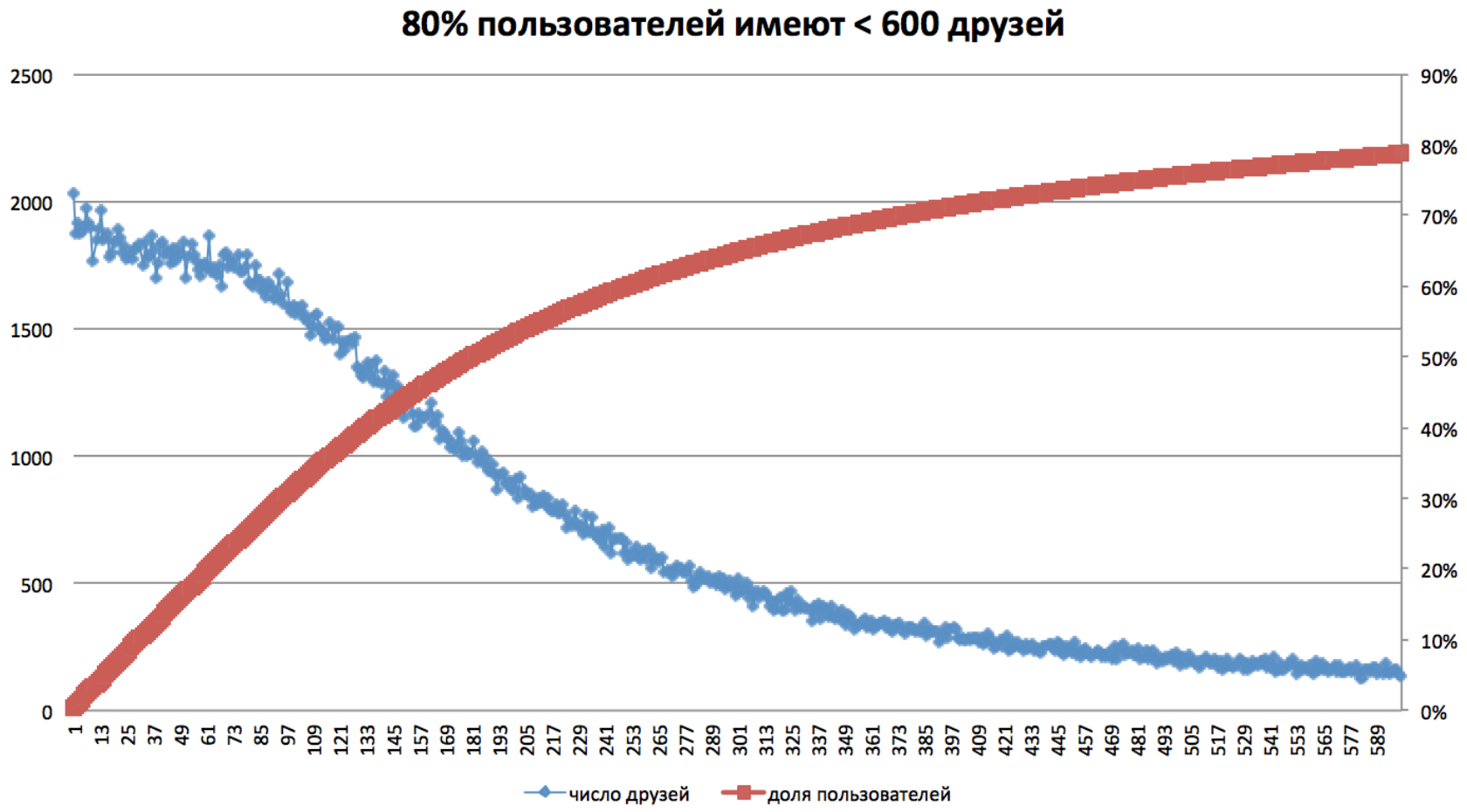
Информативность признаков

0	commonFriends.total.toDouble	332
50	avg(locationProb(demography1 friendsStats2) locationProb(demography2 friendsStats1)))	271
48	locationProb(demography1 friendsStats2) * locationProb(demography2 friendsStats1)	218
12	math.min(friendsCount1.total friendsCount2.total).toDouble	193
49	avg(genderProb(demography1 friendsStats2) genderProb(demography2 friendsStats1))	193
32	jaccard(commonFriends.total friendsCount1.total friendsCount2.total)	190
24	math.min(friendsStats1.agesMedian friendsStats2.agesMedian).toDouble	131
42	friendsCount1.schoolmate * friendsCount2.schoolmate / friendsCount1.l2norm / friendsCount2.l2norm	116
15	math.max(demography1.age demography2.age).toDouble	111
19	math.max(friendsStats1.agesMean friendsStats2.agesMean).toDouble	104
46	locationJaccard(friendsStats1 friendsStats2)	102
23	math.max(friendsStats1.agesMedian friendsStats2.agesMedian).toDouble	97
43	friendsCount1.higher * friendsCount2.higher / friendsCount1.l2norm / friendsCount2.l2norm	94
40	friendsCount1.colleague * friendsCount2.colleague / friendsCount1.l2norm / friendsCount2.l2norm	92
47	genderProb(demography1 friendsStats2) * genderProb(demography2 friendsStats1)	91
10	dotArray(countryLocation1, countryLocation2)	90
11	math.max(friendsCount1.total friendsCount2.total).toDouble	84
8	dotSparse(bigClam1 bigClam2)	83
39	friendsCosine(friendsCount1 friendsCount2)	77
14	relDiff(friendsCount1.total.toDouble friendsCount2.total.toDouble)	74

Что сработало

- Разделили пользователей на 2 группы по числу друзей
- **Число друзей < 70**: линейная модель на общих друзьях + жаккард
- **Число друзей >= 70**: уникальная модель для каждого пользователя (общих друзьях, жаккард, демография)

Число пользователей от друзей



FedorScore

$$\frac{111}{\sqrt[3]{x+50}}+8$$