1 UP 25000

2 UP 003200

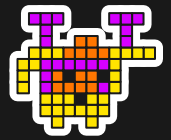
Предсказание интенсивности взаимодействия между друзьями в социальной сети ВКонтакте

Герои ML и магии

Цифровой прорыв 2023

1 UP 25000

2 UP 003200



Задача

- Предсказать интенсивность взаимодействия x между пользователями по информации об интенсивности других связей и информации о пользователях.

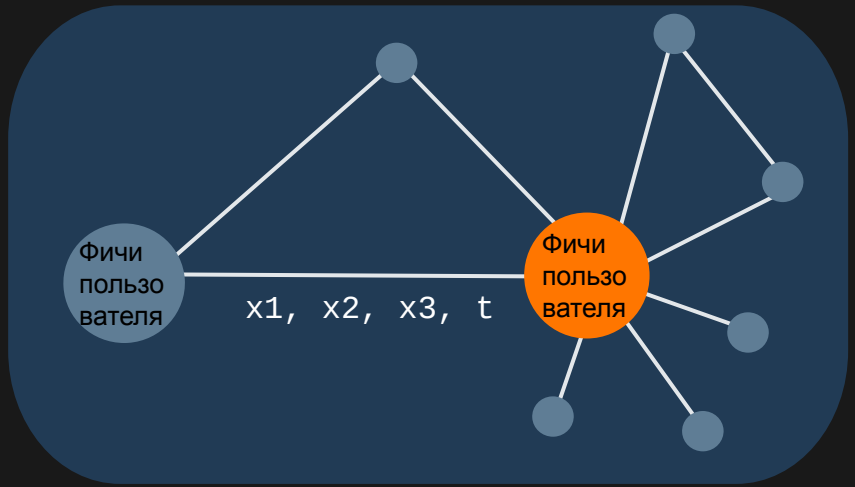
Датасет

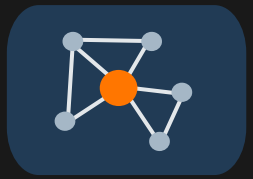
Состоит из не пересекающихся между собой эго-графов. Эго-граф состоит из центрального пользователя, его связи со всеми его друзьями, и связей между некоторыми из его друзей.

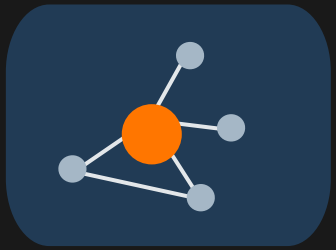
Признаки связи:
x1,x2,x3 - коэффициенты интенсивности взаимодействия
t - количество дней, которые пользователи дружат (если не дружат, то -1 (или Nan?))

Признаки пользователей:
возраст, идентификатор города, пол,идентификаторы школы и университета

Target feature:
x1

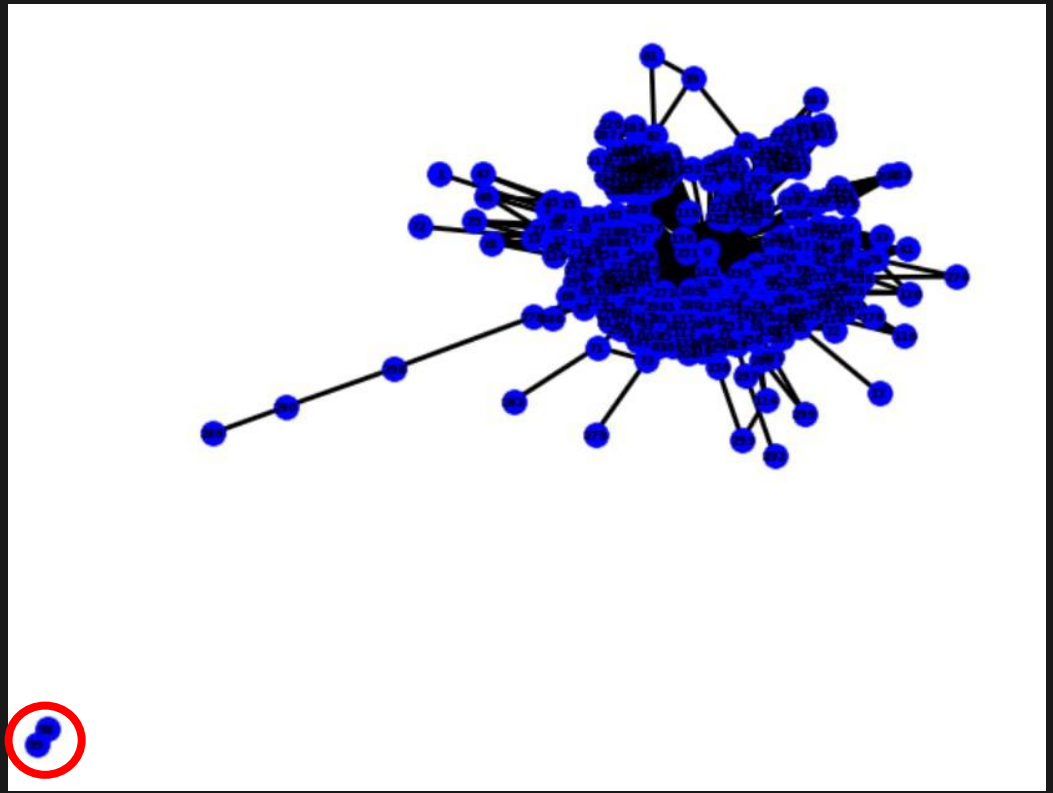






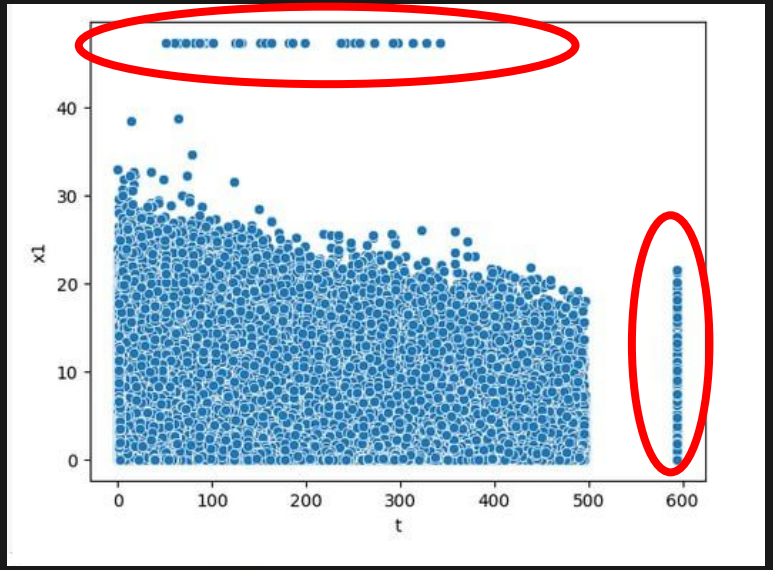
Датасет

Визуализация одного эго-графа. Видны немногочисленные выбросы в виде пользователей, не связанных с нулевым пользователем.

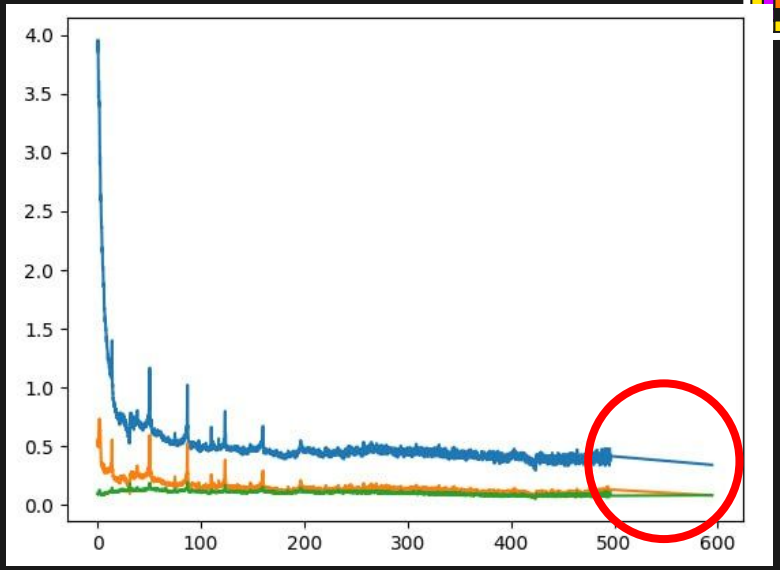


Обработка Выбросов

Записи пар друзей, имеющие x1 > 40 и t > 500, похожи на выбросы



x1(t)



x1(t), x2(t), x3(t)

Наблюдения о данных

Множества эго-графов, представленных в train и в test, не пересекаются

---

x3 – бинарный

---

x(u, v) != x(v, u) -> граф направленный

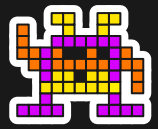
---

x2 сильно коррелирует с x1 (0.68)

Фичи пользователей, кроме возраста, – категориальные с большим количеством уникальных значений

1 UP 25000

2 UP 003200



CatBoost!

Мы сравнили CatBoost с XGBoost (1.306), RandomForests (1.330) и LinReg(XGBoost, CatBoost) (1.307), полносвязной нейросетью (1.47) и выбрали CatBoost, тк у него был лучший скор (1.306) на данных до добавления новых фичей

Генерация признаков

01

Количество общих друзей

Для каждой пары пользователей добавить количество общих друзей

02

Агрегации t

Для каждого пользователя добавить агрегацию t всех его друзей (среднее, медиана, отклонение, мин, макс)

Не дало прироста

03

Заполненность страницы

Для каждого пользователя добавить числовой показатель заполненности его страницы

Генерация признаков

04

Совпадение категориальных признаков

Для каждой пары пользователей,для каждой категориальной фичи пользователя, добавить признак, совпадает ли эта категориальная фича (0/1). Например, учились ли в одной школе.

05

One-hot encoding указанного пола

Для каждого пользователя,сделать отдельные бинарные признаки: указан пол 1, указан пол 2, пол не указан (nan)

06

Дополнительные признаки, извлекаемые из графа

-Добавить агрегации по общим друзьям u и v

-Добавить кластеризацию графов

Таблица скоров Метрика RMSE

Score ta test	Score на All Cups	Вариант модели	Фичи (по умолчанию добавлены признаки совпадения категориальных фич)
1.30	-	CatBoost	без x2, x3
1.29	-	CatBoost (после GridSearch)	без x2, x3
0.93	-	CatBoost	-
0.80	Publ 0.81 Priv 0.80	CatBoost	Агрегации, 1-hot encoding пола
0.77	Publ 0.788 Priv 0.782	**Catboost**	без агрегаций и без nan’ов, 1-hot encoding пола
1.46	-	Catboost	Без выбросов
1.47	-	Catboost	fillna(-1)

Таблица скоров Метрика RMSE

Score ta test	Score на All Cups	Вариант модели	Фичи (по умолчанию добавлены признаки совпадения категориальных фич и 1-hot encoding пола)
0.773	-	xgb	-
0.770	-	xgb	Заполненность страницы +удалены выбросы
0.84 на x3=0 и 0.76 на x3=1	-	xgb	Отдельные модели для бинарной фичи x3
0.771	-	xgb	Заполненность страницы, 1-hot encoding пола
0.758	**Publ 0.758 Priv 0.750**	**CatBoost (после GridSearch)**	**- Заполненность страницы - 1-hot encoding пола - признаки совпадения категориальных фич - количество друзей - количество общих друзей**

1 UP 25000

2 UP 003200

0.249

1-RMSE

Лучший скор (до обновления лидерборда)

Feature importance CatBoost для самого удачного решения

	feature	importance
0	t	50.206732
1	x2	32.872488
4	age_x	4.236101
7	age_y	2.682199
15	sex_y_1.0	1.488590
13	sex_x_1.0	1.472379
6	friend_cnt_x	1.435498
2	x3	1.379365
9	friend_cnt_y	1.364487
5	nan_cnt_x	1.091490
3	common_friends_cnt	1.026541
8	nan_cnt_y	0.375611
16	sex_y_2.0	0.104259
14	sex_x_2.0	0.104165
10	same_city_id	0.103073
11	same_school	0.033229
12	same_university	0.023792

Возможность работы с моделью в будущем

Генерация фичей

Не все придуманные командой фичи были реализованы по причине недостатка вычислительных мощностей и памяти. В будущем можно исследовать их

Обучение на датасете большего размера

Мы обучали модель на 10-20% данных, проверив, что расширение выборки не обеспечивает существенного повышения скора.
Но при необходимости, CatBoost легко обучить на больших данных по частям с помощью класса датасета Pool

Инференс

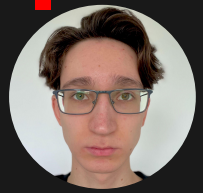
Модель мало весит и быстро работает, что позволит эффективно использовать ее для предсказания друзей пользователей на инференсе

Состав команды

Аюпова Рената

ML-разработчик

renataaupova83@gmail.com
Tg: @kinowari
Github: kinowari



Коротков Илья

ML-разработчик

i59korotkov@gmail.com
Tg: @ilkorotkov
Github: i59korotkov

Карпова Анна

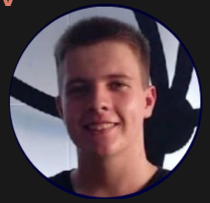
Инженер по данным

anna.a.k.2002@gmail.com
Tg: @uihlk
Github: ankkarp

Александра Куроедова

project manager, аналитик данных

sasha.kuroedova@gmail.com
Tg: @nemo926
Github: c-nemo



Булдаков Никита

ML-разработчик

nikitabuldakov@gmail.com
Tg: @BuldakovN
Github: BuldakovN

1 UP 25000

2 UP 003200

Спасибо за внимание!

Гитхаб проекта:

https://github.com/c-nemo/Edge-Prediction-on-Social-Network-Graph