ODS толпа

Дремов Дмитрий Александрович

Цель

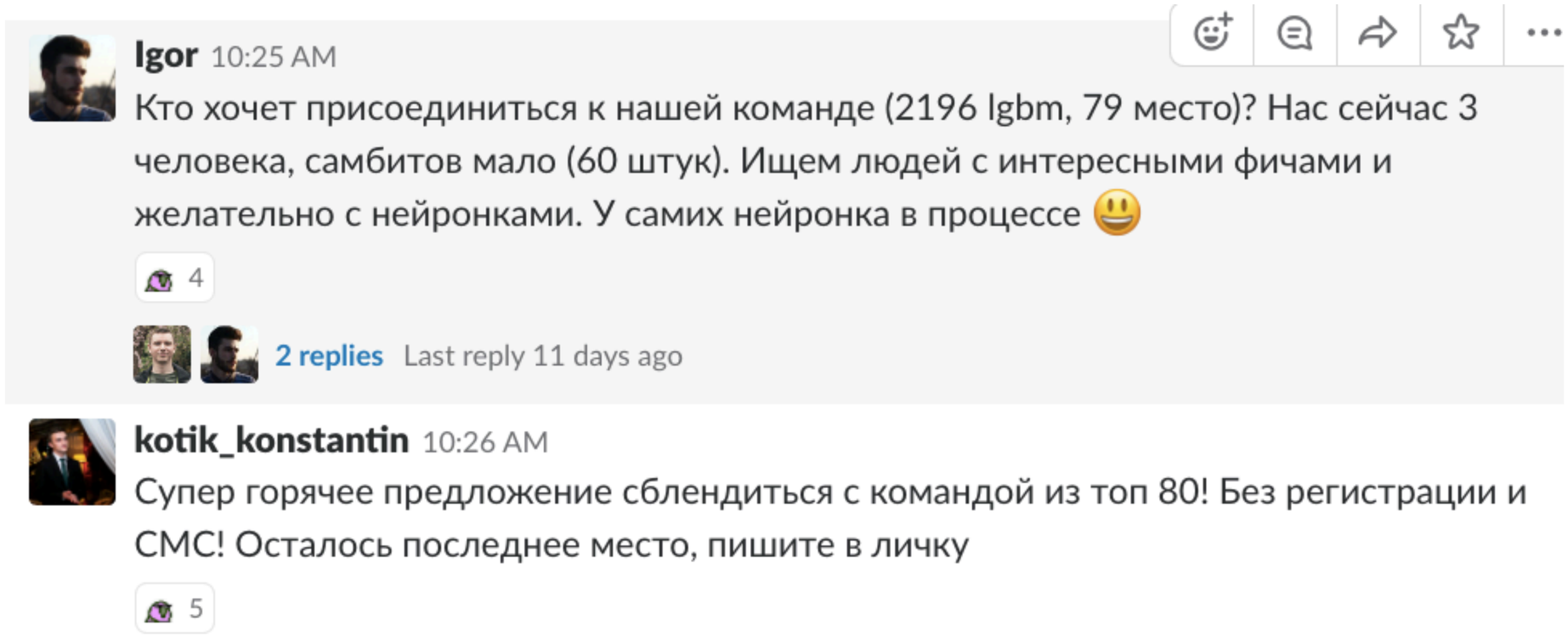
- Изначально: нахаляву получить медальки
- Итог: замарафонили в команде из 10 человек

Структура

- Как все начиналось
- Команда и взаимодействие
- Решение

Как все начиналось

Обычная история



Наша история

Tuesday, June 19th

dremovd 🐱 12:21 AM

Есть идея немного поковыряться в задачке последнюю неделю, взять паблик кернел, покрутить фичи и параметры. Возможно докинуть признаков. Ну и в итоге поблендить с какими-нибудь модельками / другими кернелами из паблика. В общем без особых претензий.

Есть существенная вероятность таким образом финишировать в бронзе / немного поднять рейтинг. Кто-нибудь хочет присоединиться к мероприятию? (edited)

➕ 1

193 replies Last reply 8 days ago

Интерес

- Всего отметилось 18 человек, необходимость организоваться
- Правила входа
- Google Docs
- Ограничения Kaggle

Организация

- Правила входа: воспроизвел любой кернел или имеет свое решение, скор не важен
- В основном самостоятельная работа
- Внутренний дедлайн за сутки. Последний день на объединение решений

dremovd 🐱 10 days ago

Я предлагаю так:
1) Входной билет в команду — иметь какое-нибудь решение. Можно быть и воспроизведенный паблик кернел, и свое решение. Скор не так важен
2) Пилить будем каждый у себя. Будем обмениваться идеями, данными, кодом и вычислительнными ресурсами по желанию.
3) За сутки до конца установим дедлайн и каждый соберет сабмит, чтобы их все заблендить. Это я могу взять на себя.

Таблица с желающими

Логин на ODS	Количество попыток	team > 3	team <= 3	В команду	Название команды	Паблик кернел в	Вычислительные ресурсы	single model	blend
TOTAL	605							2199	2189
dremovd	1	yes		ODS Crowd	ODS Crowd	https://www.kagg	i7-4770 32GB	2224	
amorgun	4	yes	yes	ODS Crowd	Alexander Morgun		i7-7700K 32Gb RAM 1x1080Ti	2241	
nv27	0	yes		ODS Crowd			Google Cloud		
KolchenkoSergey	1			??	kvigly		10x tesla P100	2246	
artgor	70			One more ODS.ai	Andrew Lukyanenko	-	i7, 64Gb RAM, 1x 1080ti	2228	2213
sharthZ23	86	yes	yes	ODS Crowd	sharthZ23	https://www.kagg	32 vCPU, 128 GB (но это не точн	2219	2211
kmike	64	yes	yes	One more ODS.ai	Mikhail Karchevskiy		6 CPU 1x 1080TI	2229	
tenich	17	yes	no	ODS Crowd	Tenich	https://www.kagg	32 vCPU, 256 GB	2241	2215
valeriy.ischenko	0	yes	yes	ODS Crowd	Valeriy Ischenko		i7 16Gb (ноут) + 40 CPU 512Gb		
leonid	16	yes	yes	ODS Crowd	Leonid Sinev	https://www.kagg	i7-2600K, 16 GB, 1080 @Win10	2238	
chslv	4	yes	yes	ODS Crowd	Oleg Chislov		Google Cloud		
Nikita	103	yes	no	One more ODS.ai	Mishunyayev Nikita	https://www.kagg	i7 8Gb (ноут)	2221	2205
vykhand	0	no	yes	eprst	eprst	https://www.kagg	Azure Cloud (plenty)		
sanbarus	0	no	yes	??	sanbarus		2680 v4, 64GB, 1080Ti		
Василий	44	yes	yes	One more ODS.ai	Looking4		ryzen 2700, 32 GB, 1070Ti	2213	2206
julia_i	0	yes		ODS Crowd	ODS Crowd	https://www.kagg	i7, 64GB, 1080	2219	
Looking4	164			eprst	eprst		i7 64GB	2199	2189
alexey_kozulin	31	yes		ODS Crowd	ODS Crowd		Azure Cloud	2243	2211

http://bit.ly/2tMN0mh

Ограничения

- Объединение не позже чем за 7 дней до финиша
- Размер команды не ограничен
- Общее количество посылок не превышает
  (5 * дней с начала = 290)
- Скор: 300/1700 соло ~ 100/1700 в команде из 5

Команда и взаимодействие

Команда

 x 1  x 4

@dremovd

@leonid

@amorgun

@nv27

@chslv

@valeriy.ischenko

@alexey_kozulin

@julia_i

@tenich

@sharthZ23

Результаты (Public LB)

День	Скор	Место	⚛	Итог	Комментарий
-7	0,2209	270		566	Объединение
-3	0,2201	170	🟤	226	Первый блендинг
-2	0,2198	140		166	+ моделей
-1	0,2193	95	⚫	110	+ моделей
0	0,2184	68		68	Лучший блендинг
0	0,2178	44		44	Стекинг
0	0,2174	35		35	Стекинг + блендинг
…экстраполяция…					
2			🟡		
7			$$$		

Взаимодействие

- Правила взаимодействия, свобода действий
- Канал в ODS для ежедневного обсуждения и координации
- Google Docs чтобы делиться идеями, информацией о моделях, сабмитах
- GitLab (LFS) + облачные хранилища для обмена кодом, признаками и предсказаниями моделей

Правила

- Сабмиты обсуждаются в чате.
  С 12 ночи по Москве можно слать что угодно
- Фиксированный holdout для train
- Фиксированные фолды для OOF

Обмен идеями

- Описание возможного признака или подхода
- Автор или источник
- Есть ли информация что идея “добавляет”
- Нужен ли GPU для работы
- Возможность “забронировать” себе идею

Объединение результатов

- Автор
- Описание модели
- Скор на holdout
- Скор на LB (если есть)
- Ссылки где лежат предсказания на holdout и test
- Для стекинга: папка в gitlab с OOF признаками

Интерес и энергичность

- Вовлечение в обсуждение
- Множество идей которые можно попробовать
- Результаты на внутреннем holdout LB
- Ежедневный прогресс скора на LB
- Помощь с конкретными задачами и сложностями
- Желание двигать команду, не быть “пассажиром”
- Магия Kaggle

Потраченное время

- За неделю несколько человек потратили 30-40 рабочих часов
- Из-за позднего дедлайна регулярно засиживались до 3-х ночи

Выводы

- Объединяться прямо перед дедлайном можно
- Большая команда может работать эффективно, но это требует аккуратной организации
- Необходимо разбиение на фолды и желательно код который с ними работает
- Важны воспроизводимые результаты с описаниями
- Полезны описания наборов признаков и желательны примеры их использования

Решение

Схема решения

- Вычисление блоков признаков
- Обучение разнообразных моделей
  train ⟶ test;
  train без holdout ⟶ holdout
- Подбор коэффициентов моделей в блендинге по holdout
- OOF генерация мета признаков
- Обучение стекинга на OOF признаках
- Линейная комбинация стекинга и блендинга

Признаки, картинки

- Размеры, соотношения, mean-std по каналам
- Neural image assessment (NIMA)
- Наличие картинки
- Предсказания ResNet

Признаки, текст

- FastText title/train
- TF-IDF по title / description
- Есть стемминг / нет стемминга
- Заголовок как категория
- Пересечение слов title и description
- Ручные фичи по тексту
- Символьные частоты

Признаки, агрегации

- Цена и логарифм цены агрегированные по группе. Средние, std, отклонение от среднего в масштабе std / в масштабе разницы 90% и 10% квантилей
- min/max/median агрегирование image_top_1 и item_seq_number по категориям
- Кодирование частотой
- Курсы валют

Модели

- LightGBM, XGBoost
- NN на sparse текстах
- Ridge на категориях, городах, регионах
- Logistic и LinearSVC на бинаризованном таргете
- Vowpal Wabbit

Трюки

- Обработка некорректных цен правилами
- Обучение отдельных моделей
  по категориям / param1
- Датасет обрезан по 2017-03-28, выкинуты записи по “3131473e84a9” и “75ebe6b373ec”

Вопросы?

Avito demand prediction: TOP-37

@dremovd

@leonid

@amorgun

@nv27

@chslv

@valeriy.ischenko

@alexey_kozulin

@julia_i

@tenich

@sharthZ23