Итоги проекта по предсказанию возраста/пола пользователя (3-е место, Д. Мошинец/Р. Цховребадзе)

Финальная конфигурация

- Фичи - просто распарсенные домены по каждому uid.
- Фичи векторизовались через T-IDF, в финальной версии пошла версия с доп. настройками
  - max_df=0.25
  - ngram_range=(4, 6)

  Почему n-граммы - жалко было терять полезную информацию, закодированную в названиях доменов, а на полноценный разбор (с обработкой транслитерации, и смысловой обработки названий доменов и т.п.) не было времени.
  Например, в случае с n-граммами если у одних пользователей есть домен driver.ru, у других - drive2.ru, то для обеих групп будет размечено и посчитано слово ‘drive’.
- Отдельная независимая классификация по полу и возрасту.
- Классификация в обоих случаях - лог. регрессия с методом newton-cg и l2-нормализацией (пришлось подкараулить свободный кластер глубокой ночью)На самом деле не очень много разницы с saga и l1-нормализации. Для модели возраста добавлен параметр multiclass=‘ovr’
- Для предсказания использовался метод predict_proba, чтобы ограничить число предсказаний на 0.5 пользователей
- Финальные модели различались способом нарезки 50% пользователей (второй способ дал прирост аж в 1% :))
  - 50% квантиль по предсказанию пола
  - Подбор отсекающей вероятности отдельно по полу, отдельно по возрасту - чтобы точно набралось 0.5 пользователей, но желательно с максимальной вероятностью

Ход решения

- Сначала домены построчно (строчки с ключом uid-домен)
- Потом перешли к исходной структуре, но с выделенными доменами (uid - массив доменов)
- Пробовали делать сразу мультиклассовую классификацию пол+возраст - на обучающей выборке accuracy была хуже, чем с отдельными предсказаниями.
- Пробовали экспериментировать с временем/днем недели - не очень зашло
- Были эксперименты с XGBoost - давал заметно лучшие результаты на исходном наборе (accuracy доходила до 0.48), лучше чем лог. регрессия (0.42). Но на проверочной выборке все поменялось

- Были эксперименты с параметрами TF-IDF, в том числе max_features - стало лучше

Что хотелось бы улучшить

- Сделать нормальную обработку IDNA-доменов
- Сделать парсинг и векторизацию самих адресов страниц (или хотя-бы векторизацию через n-граммы)
- Поэкспериментировать с вариантами моделей - последовательное предсказание пол -> возраст, сочетание разных моделей для
- Более подробно изучить что выдает векторизатор и модели - наверняка появились бы идеи как улучшить

Выводы и размышления:

- Действительно ли модель с accuracy = 0.38 на 50% пользователей лучше модели с accuracy = 0.311 на 100% пользователей?
- Очень хочется получше понять механику работу методов.
- Нужно внимательно читать набор атрибутов и методов моделей, могут появиться интересные идеи.
- Нужно значительно лучше уметь в Explarotory Data Analysis - очень хотелось получить общую картинку чего там «векторизовалось», но текущие навыки не позволяют сделать это с нужной скоростью.
- Надо учится быстро работать K-fold валидацией
- Побивать итоги - очень полезно, появляется много новых идей))