Avito Duplicate Ads Detection: обзор решений участников

Описание объявлений

https://www.kaggle.com/c/avito-duplicate-ads-detection
Avito.ru — крупный сайт объявлений.
Похожие объявления: продается одно и то же, сфотографированное с разных ракурсов или чуть-чуть по-другому описанное.
Задача: для пары объявлений предсказать вероятность того, что они являются дубликатами.

Даты: 06.05.16 – 11.07.16
Приз: $20000
Команд: 548

Данные

Набор таблиц:

- пары объявлений + целевая переменная;
- информация об объявлении: категория, геотег, текстовое описание, набор изображений, еще несколько признаков;
- архив с изображениями.

Разбиение на обучения и контроль по времени.

Итак, исходные признаки: категориальные, текстовые и графические.

Метрика и правила

Качество оценивается по AUC.
Размер тестовой выборки в ≈ 2 раза меньше размера обучающей.
Правила:

1. Объединение в команды разрешено, размер команды не ограничен.
2. Можно использовать открытые источники дополнительных данных.

Метрика

Качество оценивается по AUC.
Размер тестовой выборки в ≈ 2 раза меньше размера обучающей.
Baseline: 0.90434
Первые три места:

1. 0.95829
2. 0.95294
3. 0.94971

Генерация признаков

Участники добавляли в модель следующие признаки:

- совпадение категорий, разница в цене, число изображений, расстояние между геолокациями
- текстовые статистики: число рус. и англ. букв, цифр, длина названия и описания, число уникальных символов, расстояние cosine и jaccard между векторами 2- и 3-буквосочетаний, расстояние Левенштейна (FuzzyWuzzy) между строками-названиями и описаниями
- статистики по изображениям: (min, mean, max, std, skew, kurtosis) по каждому каналу (RGB) и среднее по трем каналам

Генерация признаков

- число незаполненных полей в каждом из объявлений
- tf-idf-признаки для текстов
- расстояния в пространстве word2vec
- число похожих опечаток в словах
- Хеши изображений (imagehash)
- расстояния между гистограммами изображений
- признаки, извлеченные утилитой imagemagick
- Dominant colour analysis

Контактная информация не помогла.

Модели

- xgboost (2500 деревьев)
- настройка xgboost - KFold
- стекинг с решениями других членов команды - Linear Regression

Код 5-го места:
`https://github.com/alexeygrigorev/avito-duplicates-kaggle`