BEST HACK

Data Science

Формулировка задания

По статистике каждый активный абонент в среднем получает более 5-ти нежелательных звонков в неделю. При этом для разных людей понятия «желаемого» и «нежелательного» трафика могут не совпадать; один и тот же номер может использоваться как для назойливого рекламного обзвона, так и для обслуживания клиентов; а на каждого нового мошенника реагировать нужно как можно быстрее.

В архиве **transactions.zip** содержатся синтетические данные транзакций голосового трафика за 2 месяца, максимально приближенные к реальным. В файле **beeline_antispam_hakaton_id_samples.csv** содержатся **id** абонентов, таргет по которым известен (train) и по которым нужно предсказать (test).

Сможете ли вы определить, к какому типу относится конкретный номер? Сможете ли вы построить стабильную и легкую модель?

Часть I - ML

Требуется построить модель, которая классифицировала бы номера на 5 категорий:

0 - не спам

1 - небольшие полезные ИП / малые бизнесы

2 - организации

3 - мобильная карусель

4 - черные спамеры и мошенники

Метрика - fbeta_score(average='macro', beta=0.5).

(https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html)

На выходе мы ожидаем файл beeline_antispam_hakaton_id_samples.csv, в котором для тестовых данных (test) проставлена одна из 5 категорий.

Часть II - R&D

Помимо точности предсказания, мы также будем оценивать применимость решения в боевых условиях и качество инсайтов, полученных из данных.

Идеальное решение должно быть не только точным, но также **стабильным** и “**легким**”.

билайн

Легким, т.к. реальные данные в тысячи раз тяжелее, а модель должна работать в режиме, приближенном к real-time, чтобы с минимальной задержкой реагировать даже на самых продвинутых мошенников, которые быстро меняют номера и подстраивают свое поведение для обхода антифрод-систем.

Под “легкостью” мы понимаем не столько сложность самого алгоритма (1 бустинг-модель вполне ОК, 10 “застеканных” моделей - нет), сколько глубину используемых данных (не обязательно использовать целый месяц, можно попробовать ограничиться одной неделей или несколькими днями, если это не сильно влияет на точность).

Под стабильностью мы понимаем постоянство качества модели в динамике как в целом по категориям, так и по каждому конкретному номеру (в идеальном мире один и тот же номер, если он не меняет фактического владельца, должен иметь одинаковый скор и вчера, и сегодня и завтра). Метрику стабильности, в особенности для случая с конкретными номерами, мы предлагаем вам придумать самим.

В частности, нам бы хотелось, чтобы вы постарались ответить на следующие вопросы (но, возможно, вы захотите поисследовать в данных что-то еще):

1. Какие паттерны поведения номеров помогают отделить категории 1-4 друг от друга?
2. Насколько отличаются предсказания модели, построенные на разных периодах времени?
3. Не “ломается” ли модель в выходные и праздники?
4. Какое минимальное количество дней исторических транзакций требуется, чтобы построить стабильную и достаточно точную модель?
5. Какую метрику стабильности скора по отдельным номерам вы предлагаете использовать?
6. Как модель работает на слабоактивных номерах?
7. Среди номеров из наиболее “спорных” категорий 2 и 3, возможно ли выделить полезный и нежелательный трафик на уровне конкретного звонка?

Мы очень приветствуем, чтобы ваши выводы были подкреплены конкретными цифрами и графиками (визуализация должна подтверждать тезис и быть читаемой, но её красота сама по себе не оценивается).

Удачи!

Датасет вы можете найти по ссылке:
https://drive.google.com/drive/folders/1Bn3bvV5u15a7enelJwmqyRrWu9Dfm-eZ?usp=sharing

билайн

Описание данных

- id_a - id абонента, который звонит
- id_b - id абонента, которому звонят
- time_key - дата звонка
- start_time_local - время начала звонка
- time_zone - часовая зона звонящего абонента
- duration - длительность звонка
- forward - индикатор переадресации
- zero_call_flg - категория звонка с нулевой длительностью
- source_b - индикатор транзакции из источника B
- source_f - индикатор транзакции из источника F
- num_b_length - длина номера абонента, которому звонят

Требования к решению:

Перед отправкой решения убедитесь, что оно соответствует следующим критериям:

- Присутствует презентация вашего решения в формате .pptx или .pdf, в которой присутствует следующая информация:
  - Название команды;
  - Имя капитана команды;
  - ФИО членов команды;
  - Подробное текстовое описание решения с обоснованием выполненных действий;
  - Визуализация данных и работы модели;
  - Результаты, полученные при вычислении указанной метрики качества.
- Присутствует файл (ноутбук) с программой в формате .ipynb (Jupyter Notebook, Google Collab, Kaggle Notebook и т.п.), где подробно описано решение. Обратите внимание, что требуется именно файл, а не ссылка на него.
- Файл **beeline_antispam_hakaton_id_samples.csv**, в котором для тестовых данных (test) проставлена одна из 5 категорий

Решение отправляйте на почту besthackathon2022@gmail.com с темой **[Ф] [Data Science] Название команды**. В теле письма обязательно должны быть указаны ФИО капитана команды.

билайн