### **Домашнее задание 2**

Основная задача - по данным постов в телеграм каналах предсказать количество просмотров этих постов.

Поскольку мы изучаем Spark, акцент в задании - на подготовку фичей на кластере, а не на непосредственно тюнинге модели. Проверяться качество модели тоже будет, но достаточно будет преодолеть довольно простой бейзлайн.

[Данные](https://yadi.sk/d/iwvBZBEbSxrxKA)

Вот тут лежит шаблон решения <https://yadi.sk/d/GBQVSv4VpUdDEA>

Там предлагается лишь один из возможных путей решения, можно модифицировать его как угодно.

#### **Данные.**

Предоставляется 3 файла:

1. трейн сет (с количеством просмотров)
2. тест сет (без количества просмотров)
3. метаданные каналов

Трейн и тест разбиты по времени. Как прочитать данные можно посмотреть в ноутбуке.

#### **Задание.**

1. Подсчитать фичи для модели, используя только Spark
2. Фичи можно перевести в pandas и обучить свой любимый алгоритм локально (но не обязательно)
3. Предсказать им тест сет и побить бейзлайн по целевым метрикам

#### **Метрики.**

Поскольку просмотры распределены экспоненциально, предсказывать будем странную величину log(post\_views + 100). Вычисляются сразу 4 метрики, но они связаны между собой. Это

1. RMSPE - Root Mean Squared Percentage Error
2. MAPE - Mean Absolute Percentage Error
3. MAE - Mean Absolute Error
4. RMSE - Root Mean Squared Error

#### **Сабмиты.**

Для проверки сабмитов используется специальный модуль. Как его использовать смотрите в ноутбуке.

При импорте этого модуля у вас запросит имя пользователя и пароль - их мы вышлем всем лично.

Если формат предсказаний неправильный, вернется ошибка с описанием. Если все хорошо, вернутся значения метрик.

**За день можно делать максимум 5 сабмитов. Для вашего итогового сабмита нужно выставить флаг final=True для функции make\_eval. Это сохраняет ваш сабмит для проверки, без этого задание считается не сделанным!**

#### **Критерии.**

Для получения оценки > 0 нужно прислать ноутбук в виде .ipynb файла, который можно выполнить через Run All, и который в результате сделает сабмит. Прислать нужно на почту [mike0sv@gmail.com](mailto:mike0sv@gmail.com)

Для получения положительной оценки (>3) нужно сделать сабмит с метриками лучше, чем у бейзлайна

BASELINE = {

'mape': 15.707128974856676,

'mean\_absolute\_error': 1.219070382113261,

'mean\_squared\_error': 2.4324378881170055,

'rmse': 1.5596274837655963,

'rmspe': 23.50065988751091

}

Выполнение этого условия дает вам оценку 5.

Остальные баллы можно получить за следующие пункты.

1. 1 балл за фичу, использующую метаданные каналов (макс 1 балл)
2. 2 балла за фичу, использующую текстовые поля (макс 2 балла)
3. 1 балл за фичу, использующую window aggregation, до 2 штук, различающихся по смыслу и реализации. Фичи которые можно альтернативно подсчитать групбаем и джойном не засчитываются (макс 2 балла)
4. -1 балла за грязный код и -1 за отсутствие комментариев
5. -10 баллов за списывание))) (обоим участникам) (макс -10 баллов)

Все фичи должны быть осмысленными для задачи. Поля исходных данных и их простейшие преобразования (математические выражения для числовых, количество символов/слов для текстовых и тд) не засчитываются. Но при этом вы можете их добавлять чтобы улучшить модель

FAQ:

* Что означает признак tg\_id в данных?
  + внутренний айдишник телеграма
  + если речь про посты, то это ~номер поста в канале, то есть они уникальные в рамках одного канала
* что такое last\_parsed в channel\_data? что такое updated в channel\_data? зачем нужен id?
  + это остатки системы для кравлинга. даты использовались для определнеия когда какой канал нужно перепарсить, айди - из бд где хранились данные
* Ошибка при отправке сабмита
  + проверьте что ваши df.values имеет такой же тип и шейп как np.zeros(testX.shape[0])