Входные данные

Файл **train.csv** содержит набор твитов, доступных для обучения. Каждая строка файла соответствует одному твиттер-сообщению с дополнительной информацией о нем и его авторе. Первая строка файла содержит заголовок таблицы с названиями колонок.

Файл **test.csv** имеет полностью аналогичную структуру, за исключением отсутствия последней колонки *retweet_count*.

Вам необходимо разработать модель, которая будет предсказывать вероятность того, что значение в последней колонке ***retweet_count*** будет больше 20, используя данные из остальных колонок.

Выходные данные

Результатом работы Вашей программы должен быть файл **prediction.csv**. Он должен содержать ровно 2 колонки. В каждой строке через запятую должны быть записаны *id* сообщения из **test.csv** и вероятность того, что *retweet_count* сообщения больше 20. Первая строка файла — заголовок таблицы.

Внимание!

Твиты в обучающем и тестовом наборах данных написаны непересекающимися множествами авторов, потому что мы хотим, чтобы Вы в рамках конкурса попытались построить модель с несколько большей степенью обобщения, чем элементарное заучивание уникальных идентификаторов авторов. **Учитывайте это при построении и валидации модели! Качество модели будет оцениваться площадью под ROC-кривой, построенной на основе прогноза на тестовом наборе данных.**

Взаимодействие с сетью Интернет вашего решения категорически запрещается. Внешние данные можно использовать только по предварительному согласованию с Организаторами конкурса.

Обращаем ваше внимание, что при дополнительной проверке вашей модели ваш код может быть запущен на других входных данных, собранных по аналогичной методике.

Пример решения

Вам предлагается пример решения **Demo.ipynb** в формате IPython Notebook, который демонстрирует чтение входных данных, построение простейшей модели и запись в выходной файл. Кросс-валидации модели в примере нет, вам придется ее добавить в код самостоятельно. Для удобства дополнительно доступны экспорты примера **Demo.py** и **Demo.html**.

Что вы должны прислать в качестве решения?

1. Прогноз **prediction.csv** для твитов из файла **test.csv**.
2. Код программы рекомендуется присылать в формате IPython Notebook (в файле Solution.ipynb). Результат полного выполнения вашей программы — файл **prediction.csv** .
3. Отчет в формате PDF о проведенном исследовании. Изложите в нем основные использованные идеи, ключевые особенности вашего алгоритма. Допускается предоставление отчета в формате IPython Notebook вместе с кодом.
4. Если Вы предоставляется решение в формате, отличном от IPython Notebook, обязательно напишите в отчете, как его запустить.

Любые вопросы и ваши решения присылайте до 11:59:59 9 января включительно на электронный адрес lab@indatalabs.com . Успехов!