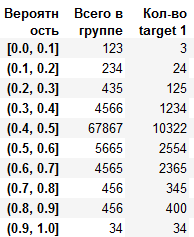
Даны два файла **data.csv** и **verify.csv**. Разделитель – ';'

Поля: ID – уникальный айди записи, feature\_1 … feature\_124 – признаки.

В файле data.csv также присутствует **target**, принимающий два значения – 0 или 1

1. **Задание**
   1. Решите задачу двухклассовой классификации, обучив модель на данных в **data.csv** После обучения вы должны получить вероятность принадлежности каждого объекта к таргет, равному единице 1.
   2. Сгруппируйте выборку по значениям вероятности с шагом 0.1, и в каждой группе посчитайте: сколько объектов в группе, сколько в группе объектов с target = 1.

Например так:

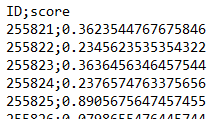


* 1. Посчитайте метрики модели (F-мера, ROC-AUC, коэффициент Джини и другие) на данных из data.csv

Основная метрика для задания – **ROC-AUC**.

* 1. С помощью полученной модели определите вероятность принадлежности к таргет 1 каждого объекта из **verify.csv.**
  2. Запишите полученные предсказания в csv-файл (разделитель – точка с запятой). В файле должны быть два поля: ID – айди который задан в файле **verify.csv** и score – вероятность которую вы получили для данного ID.

Пример файла:



Таким образом, по результатам задания вы должны получить:

* метрики модели на данных из data.csv
* таблицу с группировкой
* файл с предсказаниями на данных из файла verify.csv
* записки/текстовые файлы с кратким описанием хода обучения

Необходимо использовать **логистическую регрессию.** Если есть желание, дополнительно можете прислать результаты любой другой модели, если считаете, что она лучше.

1. **Вопросы**

Ответьте на следующие вопросы:

* Как можно заполнить пропущенные значения в данных?
* Что делать, если таргет несбалансированный?
* Как бороться с переобучением?
* Найдите потенциальные причины аномалий в наборе данных.
* Что делать, если данных не хватает или они плохого качества?
* Что можно сделать чтобы повысить качество?