**Задание: Предсказание успеха песни**

В этой задаче мы построим модель для предсказания успеха песни. Наша цель

предсказать, попадет ли песня в список TOP10. Это задача классификации – попадет\не

попадет. Мы решим ее с помощью логистической регрессии.

1. Скачиваем данные по ссылке:

<https://drive.google.com/file/d/1x3CwvM0d_ok-WBmtUtZEICHM2U4nOVN8/view>

и добавляем их в Orange

1. О датасете

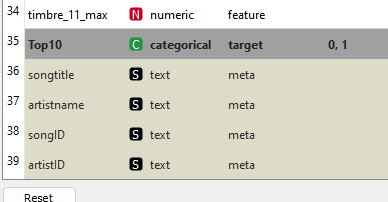
Датасет хранит в себе 7574 строк и 35 переменных без пропущенных значений

Переменные датасета:

1. year = год выпуска песни
2. timesignature и timesignature\_confidence = переменная, оценивающая временную подпись песни ( это соглашение в [западной](https://en.wikipedia.org/wiki/Western_culture" \o "Западная культура) [музыкальной нотации](https://en.wikipedia.org/wiki/Musical_notation), которое определяет, сколько [значений нот](https://en.wikipedia.org/wiki/Note_value" \o "Значение примечания) определенного типа содержится в каждом такте), и уверенность в этой оценке
3. loudness = непрерывная переменная, указывающая среднюю амплитуду звука в децибелах
4. tempo и tempo\_confidence = переменная, указывающая на количество ударов в минуту в песне, и уверенность в этой оценке
5. key и key\_confidence = переменная с двенадцатью уровнями, указывающая предполагаемую тональность композиции (C, C #, . . . , B) и достоверность оценки
6. energy = переменная, представляющая общую акустическую энергию песни с использованием набора функций, таких как громкость
7. pitch = непрерывная переменная, указывающая высоту звука песни
8. timbre\_0\_min, timbre\_0\_max, timbre\_1\_min, timbre\_1\_max,…, timbre\_11\_min и timbre\_11\_max = переменные, указывающие минимальные / максимальные значения во всех сегментах для каждого из двенадцати значений в векторе тембра (в результате получается 24 непрерывных переменных)
9. Top10 = двоичная переменная, указывающая, попала ли песня в топ-10 чарта Billboard Hot 100 (1, если она была в топ-10, и 0, если нет)
10. songtitle = название песни
11. artistname = имя исполнителя песни
12. songID и artistID = определение переменных для песни и исполнителя
13. Перед тем как анализировть данные, помечаем в Orange переменную Top10 как нашу цель для исследования (target), а переменные year, songtitle, artistname, songID и artistID как метаданные (meta) – данные о данных, не используемые для статистических выводов.

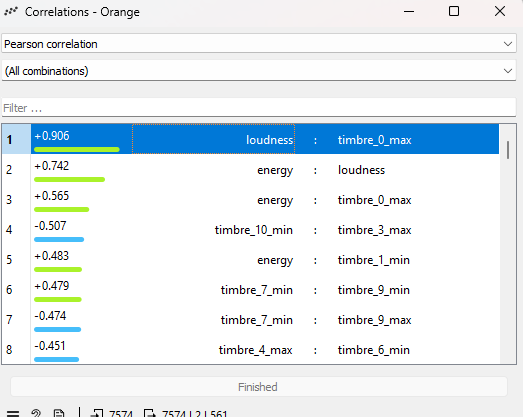
A screenshot of a computer

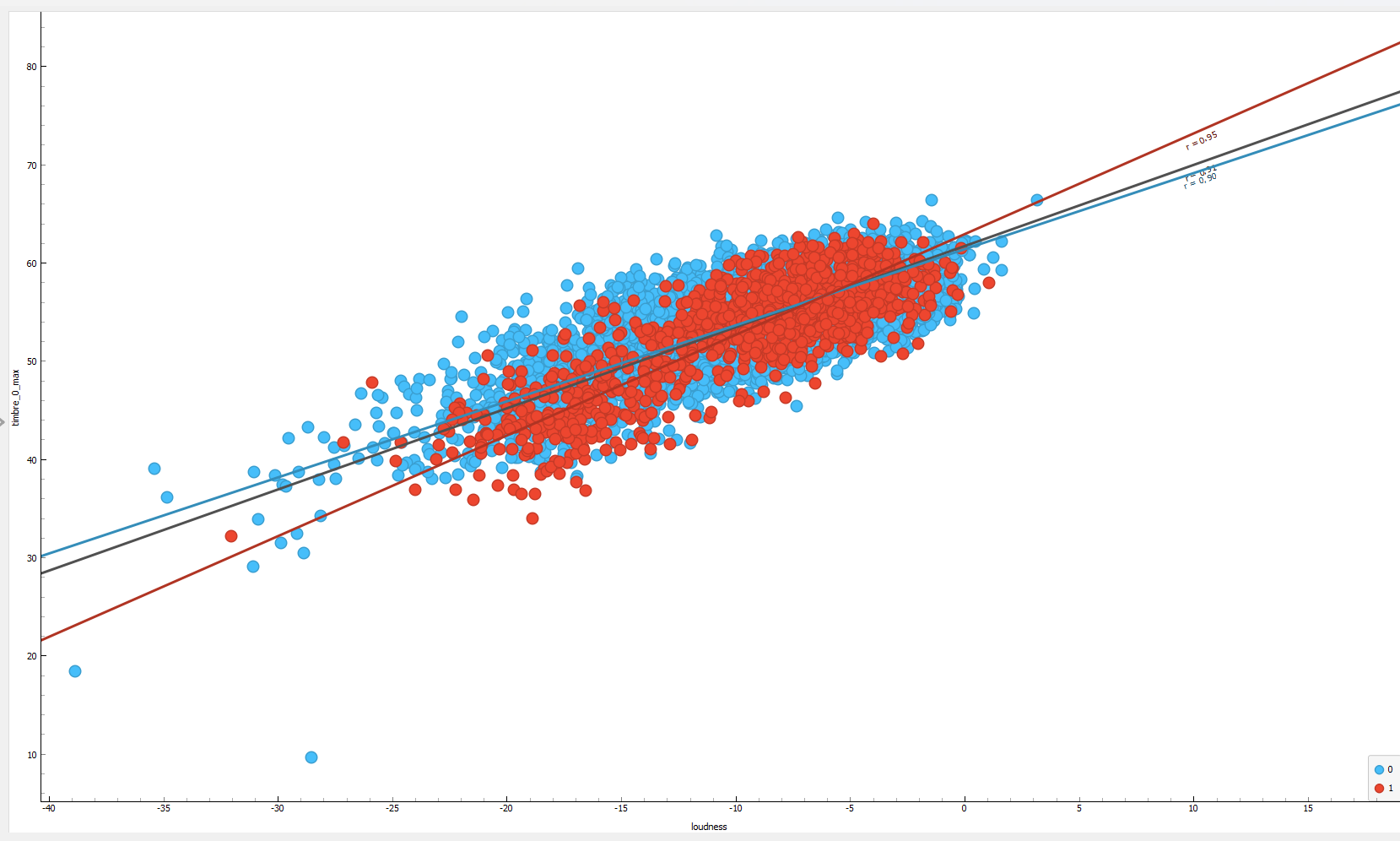
Description automatically generated



1. Проводим разведывательный анализ данных

С помощью модуля Correlations мы можем заметить, что есть большая положительная корреляция между двумя независимыми параметрами loudness и timbre\_0\_max, что может создавать мультиколлинеарность логистической регрессии.

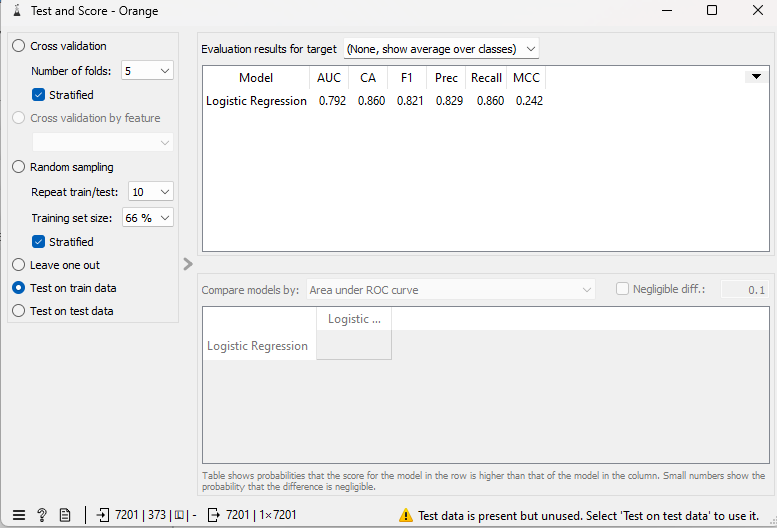




Поэтому, чтобы избежать дальнейших проблем, можно убрать переменную timbre\_0\_max (feature -> skip)

1. Построение и оценка модели

Для создания модели мы разделим данные по принципу песни до 2009 года (включительно) и песни, вышедшие после 2009. Первая группа пойдет в тренировочные данные, вторая-в тестовые.

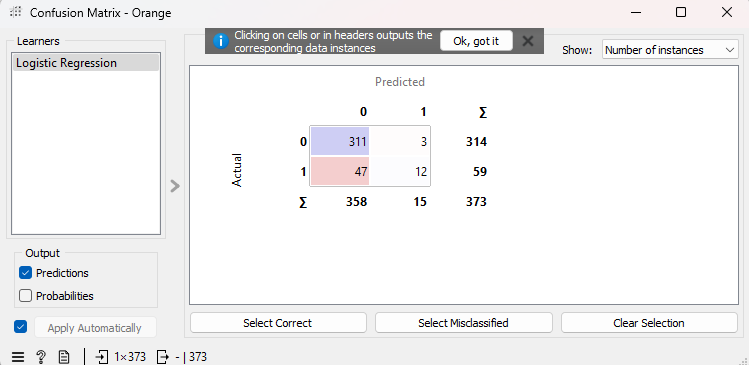


A screenshot of a computer

Description automatically generated

Тест на тренировочных данных показал значение AUC 0.792, а на тестовых – 0.833. Это говорит нам о том, что наша модель достаточно хороша для прогнозирования попадания песни в топ-10.

Присоединяем к модулю Test and Score модуль Confusion Matrix



Данная матрица говорит нам о том, что присутствует ошибка второго рода. То есть, наша модель препдолагает, что песня не войдет в топ-10, однако на самом деле она входит. Если бы была ошибка первого рода, то модель преполагала бы, что песня войдет в топ-10, однако на самом деле нет. Что хуже? Предположительно – второго рода, так как мы теряем большую возможность попасть в топ-10. Также с помощью этой матрицы можно понять, что чаще модель ошибается на данных из топа.

С помощью модуля Nomogram, можно увидеть что тремя основными факторами, предсказывающими успех песни являются: высота звуа песни (pitch) (чем выше – тем ниже успех), минимальное значение в 11 значении вектора тембра (timbre\_11\_min) (чем выше- тем лучше) и минимальное значение в 1 значении вектора тембра (timbre\_1\_min) (чем выше – тем лучше)

