НЕТОЛОГИЯ

Юстина Иванова

Программист, data scientist

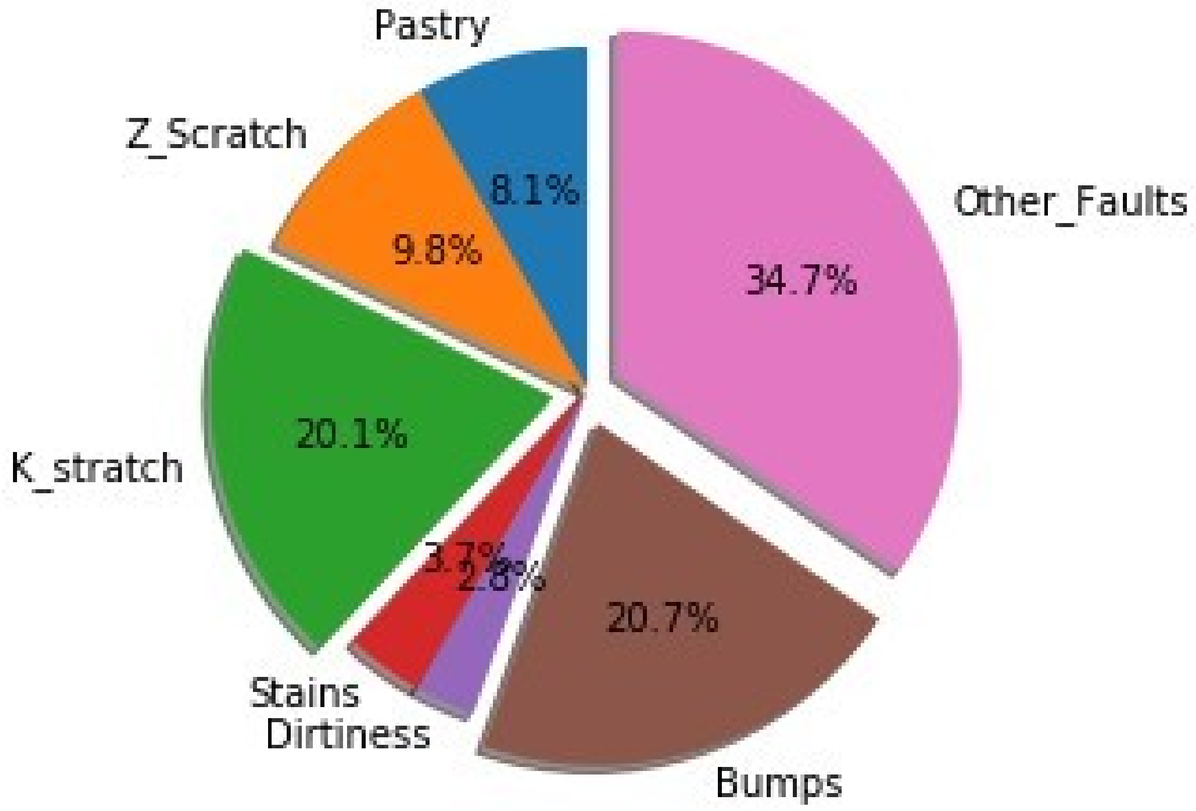
Статистика в python. Кейс-стади №1.
Датасеты: faulty steel plates,
heart disease record,
Brent oil prices.

Спикер

Юстина Иванова,

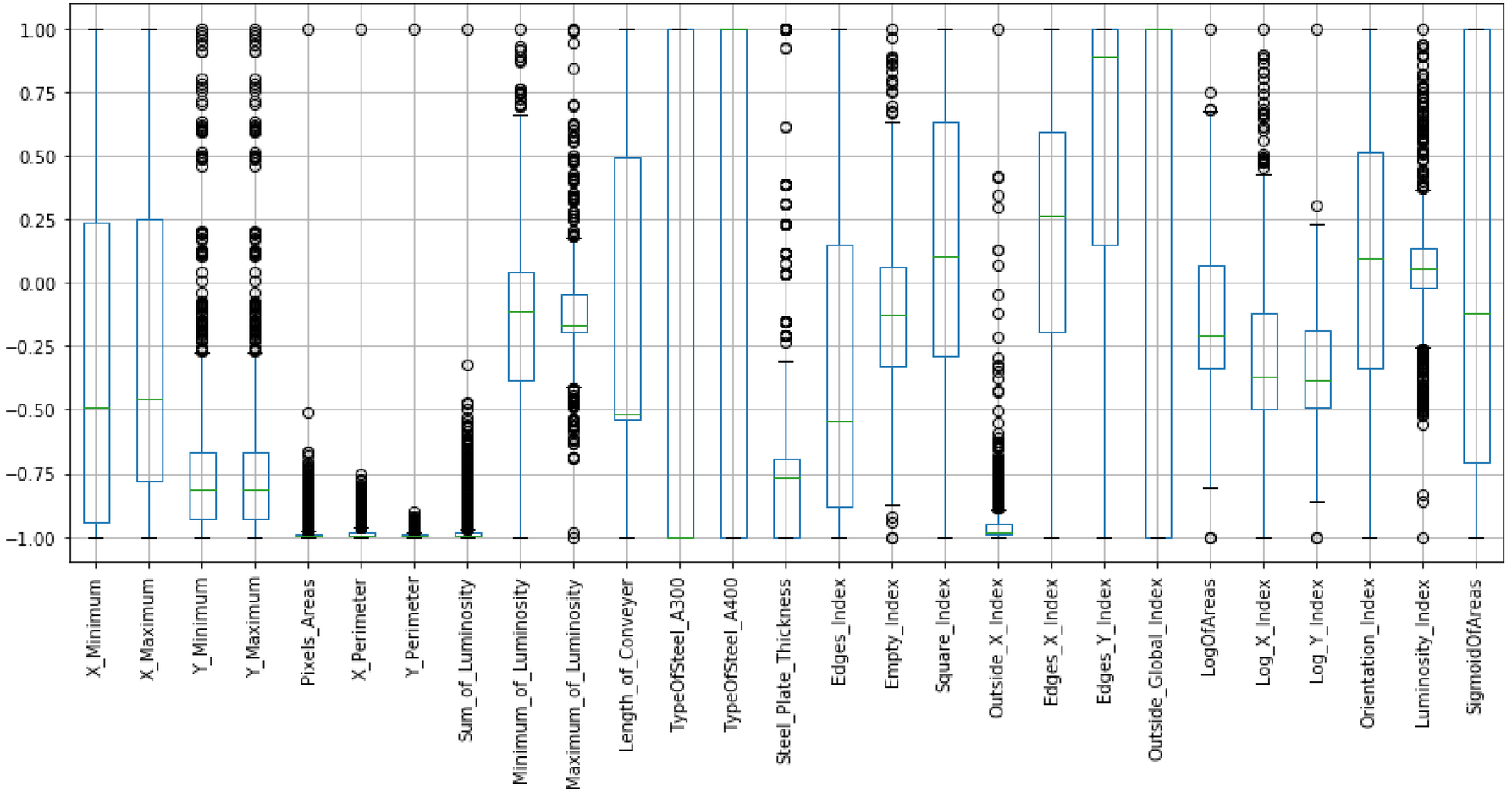
- PhD в Унивеситете Больцано
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence в University of Southampton

Датасет Faulty Steel Plates.



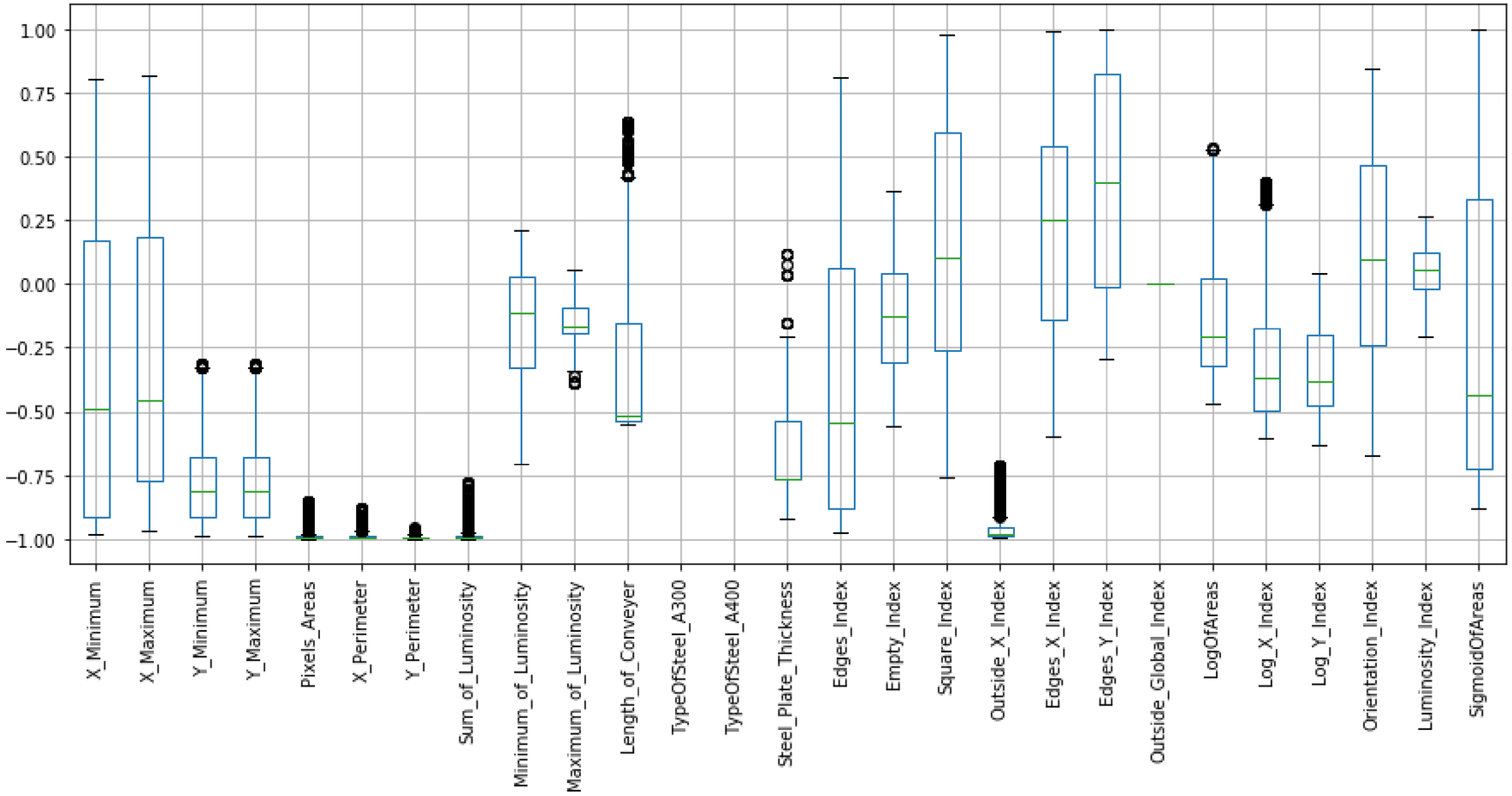
https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

Boxplot → выбросы.



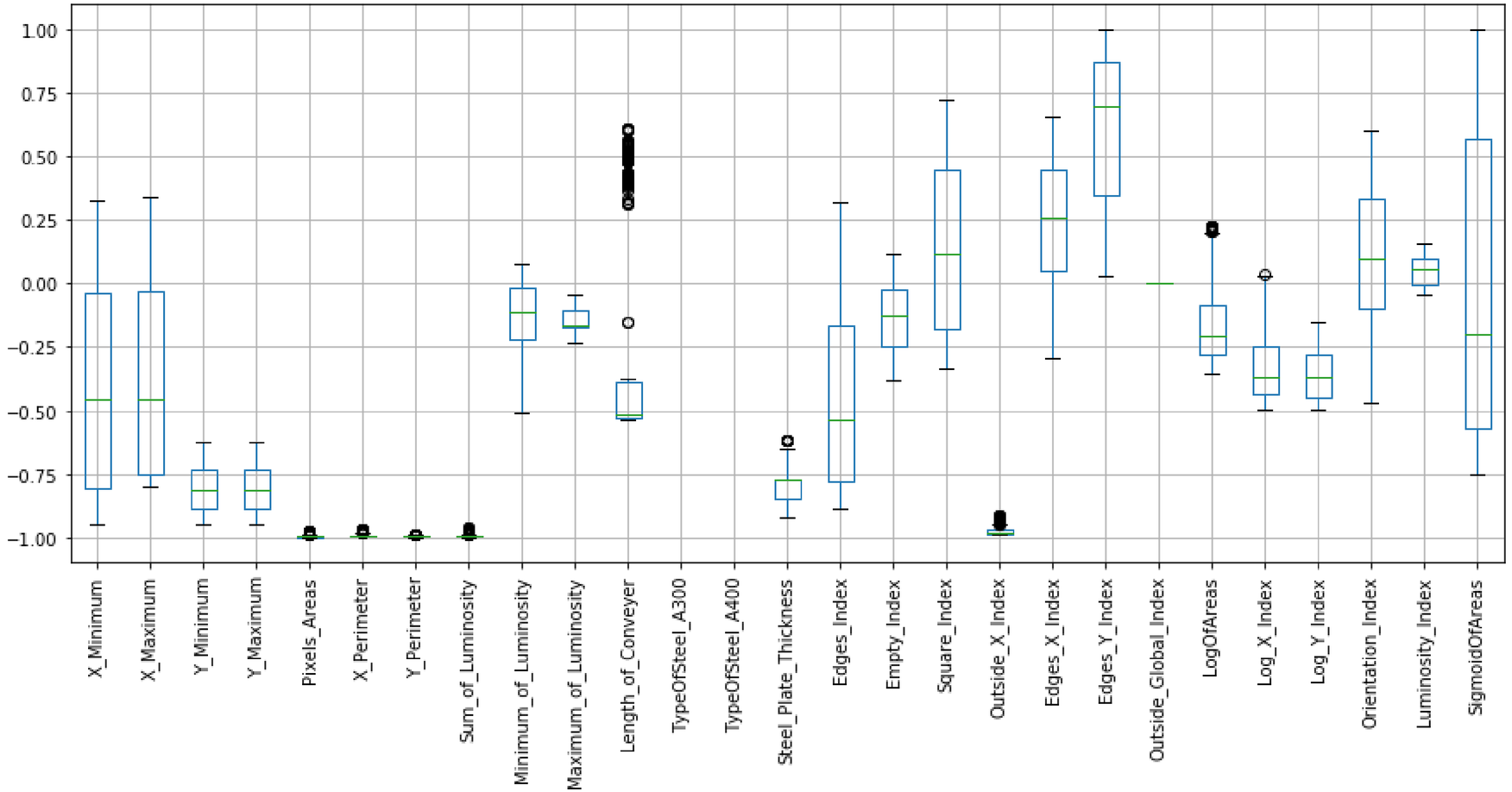
https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.boxplot.html

Удаление элементов вне интерквартильного интервала..



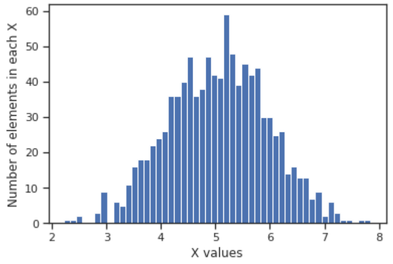
https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.boxplot.html

Удаление элементов вне квантилей 20% и 80%.

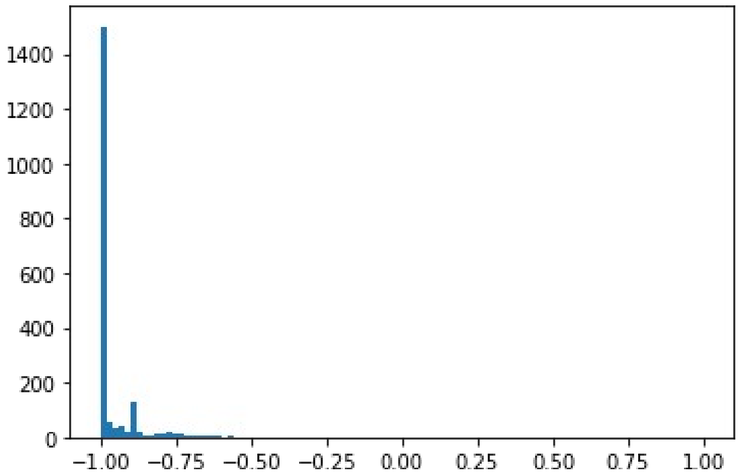


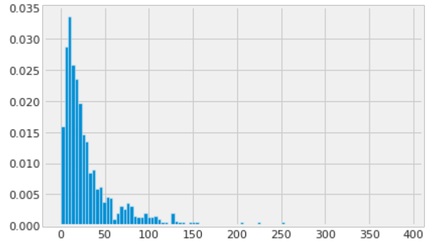
https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.boxplot.html

Тесты на согласие: какое это распределение?

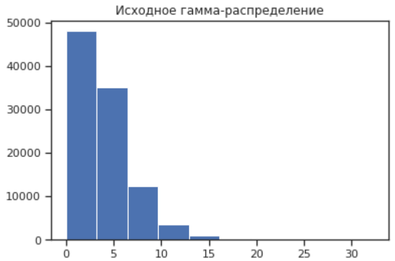


нормальное

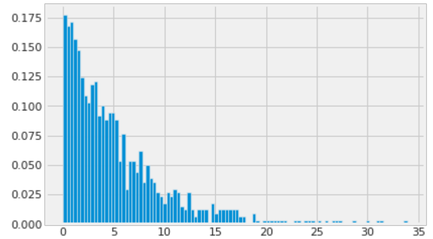




экспоненциальное



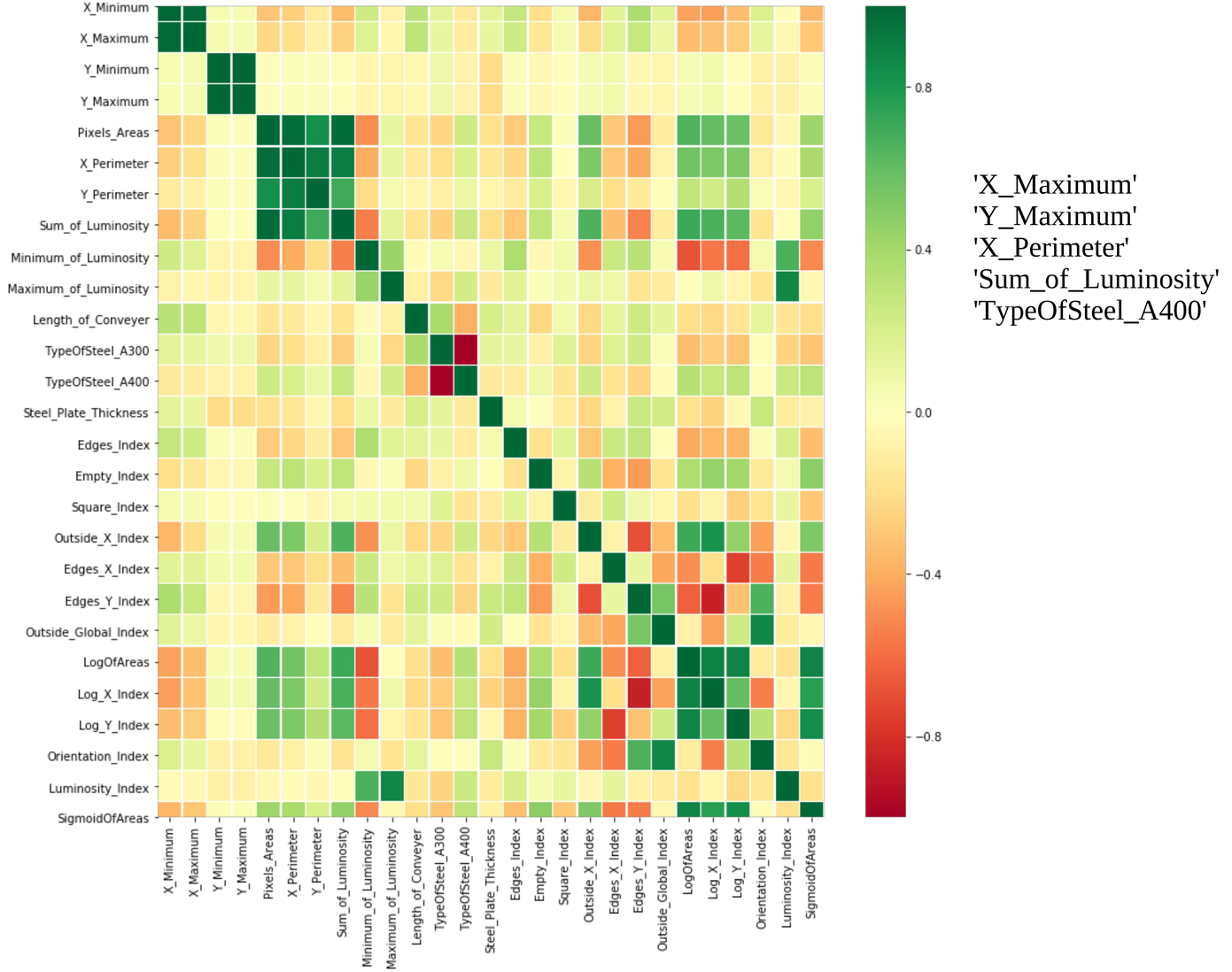
гамма



экспоненциальное

https://mc.ai/15-statistical-hypothesis-tests-in-python-cheat-sheet/

Удаление мультиколлинеарности



'X_Maximum'
'Y_Maximum'
'X_Perimeter'
'Sum_of_Luminosity'
'TypeOfSteel_A400'

https://ru.wikipedia.org/wiki/Снижение_размерности

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Классификация

«Разложи носки по цвету»

Предсказать значение

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Кластеризация

«Разложи похожие вещи по кучкам»

Найти зависимости

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Выявить последовательности

Ассоциация

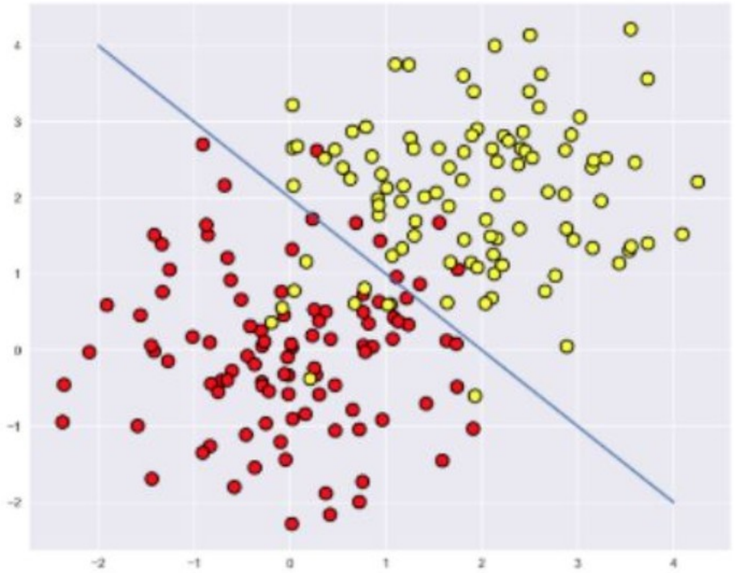
«Найди какие вещи я часто ношу вместе»

https://vas3k.ru/blog/machine_learning/

НЕТОЛОГИЯ

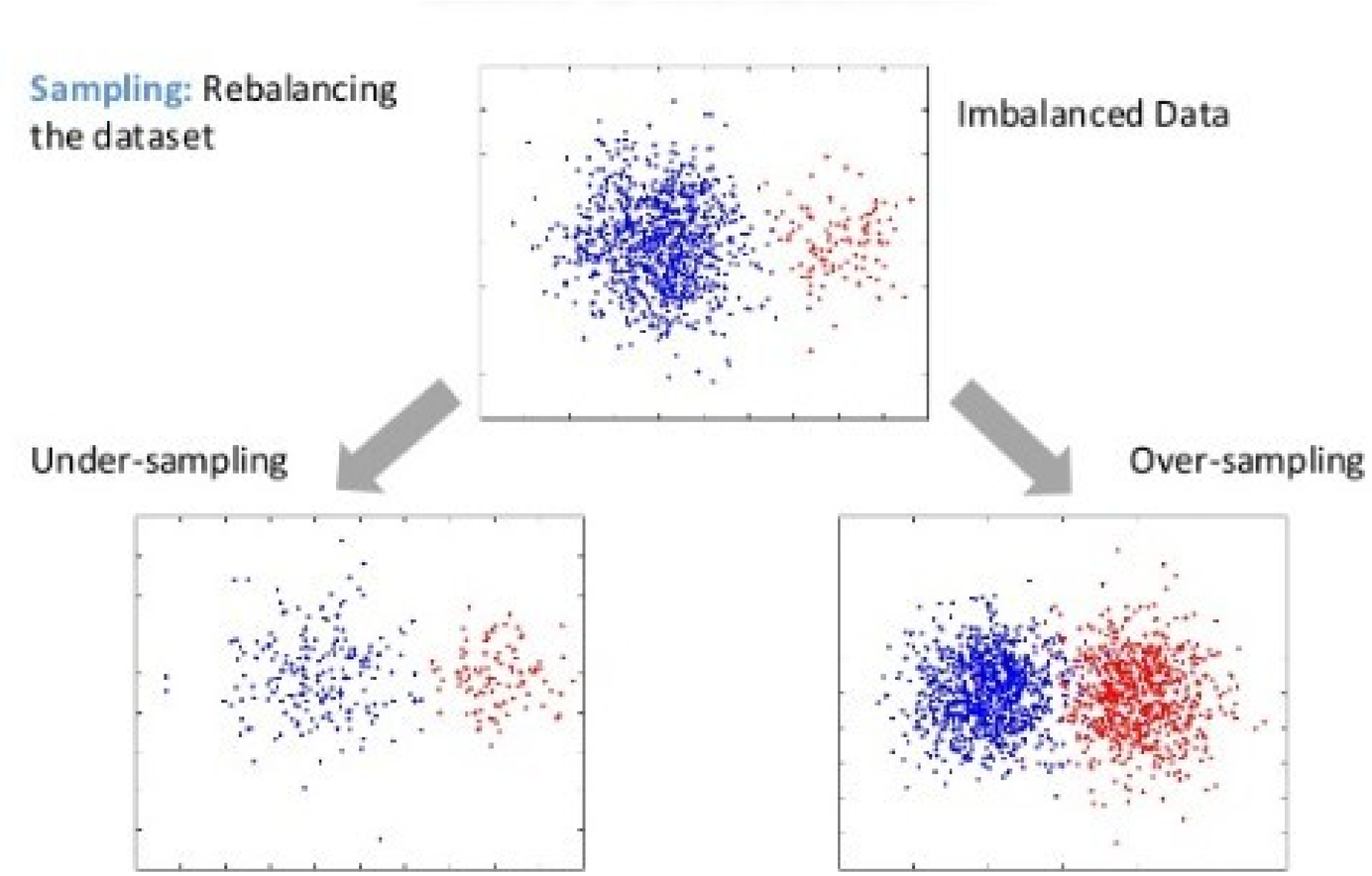
Классификация

Множество допустимых ответов конечно. Их называют метками классов (class label). Класс — это множество всех объектов с данным значением метки.



https://github.com/miptmlschool/

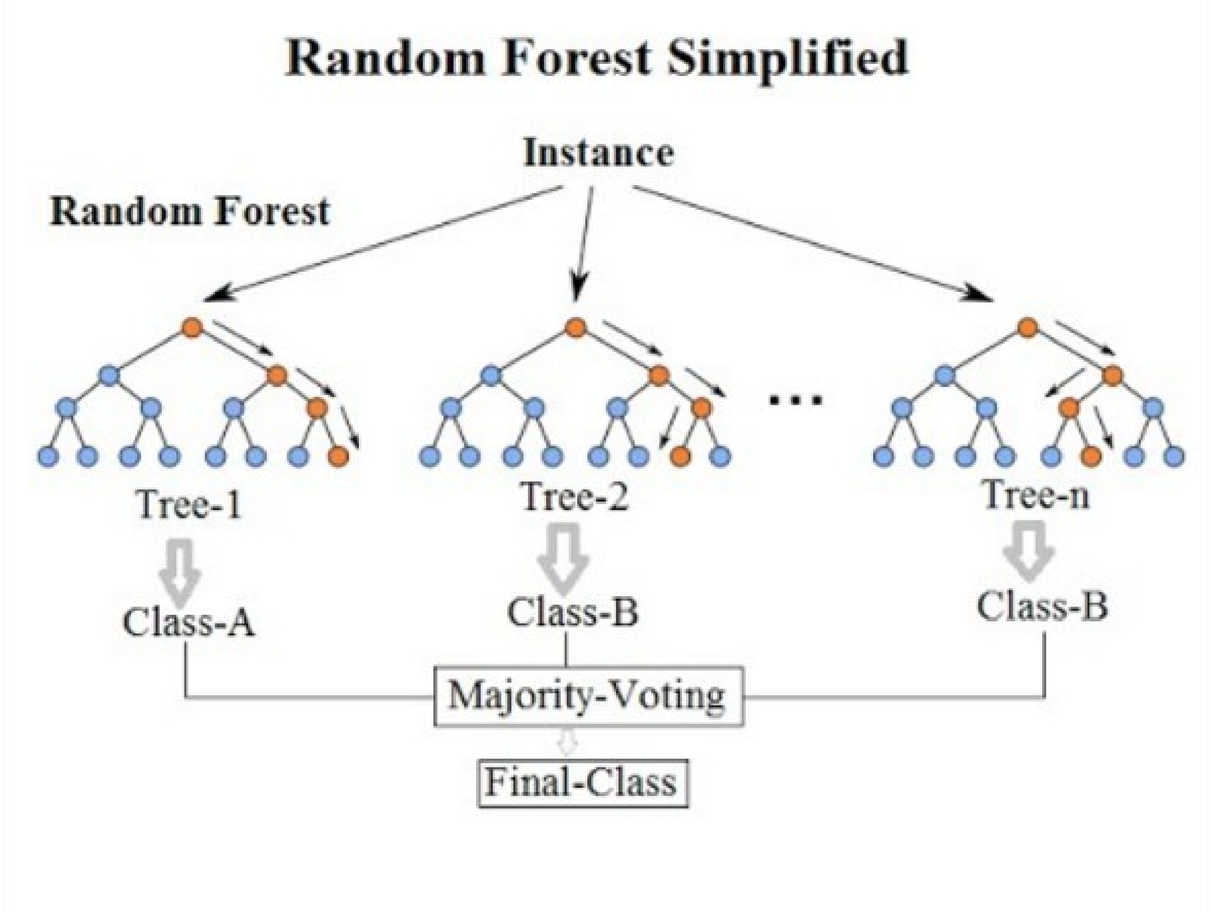
Проблема несбалансированности классов.



https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family

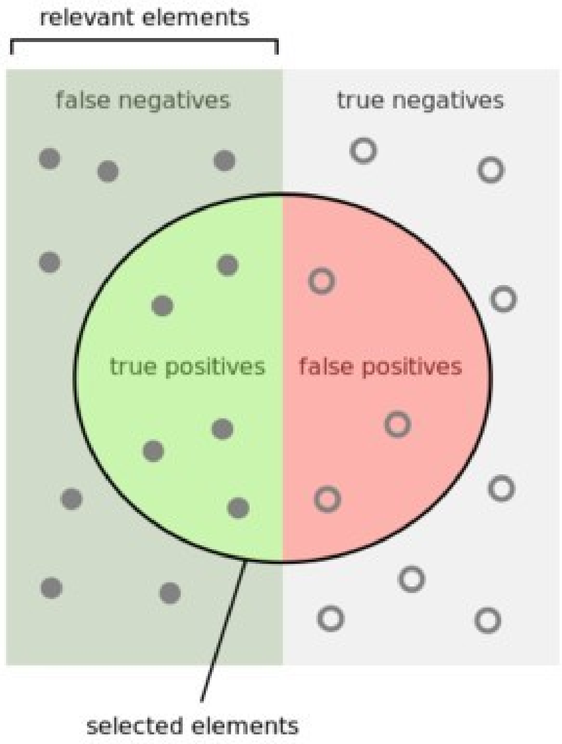
Дерево решений.

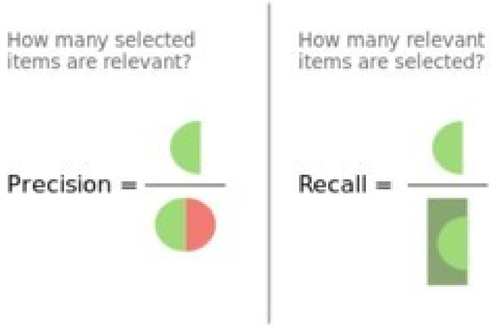
Случайный лес.



https://ru.wikipedia.org/wiki/Random_forest

Метрики классификации





Precision

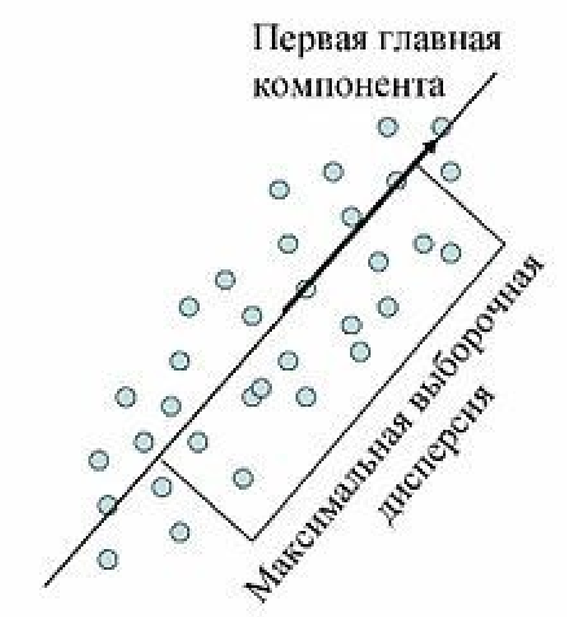
Recall

F1-мерамера

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

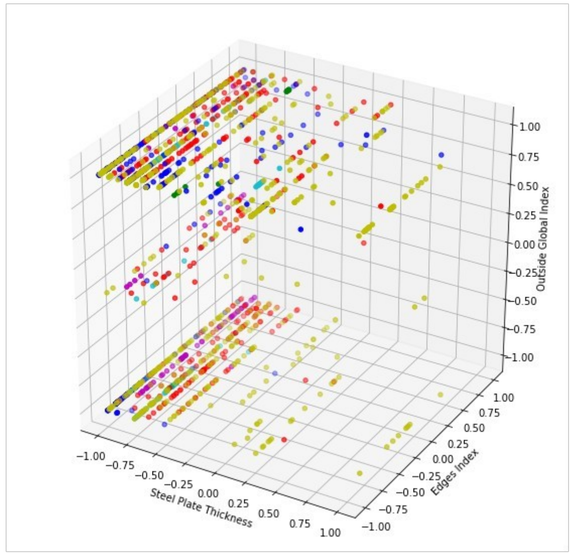
https://habr.com/ru/company/ods/blog/328372/

Принцип минимальных компонент.

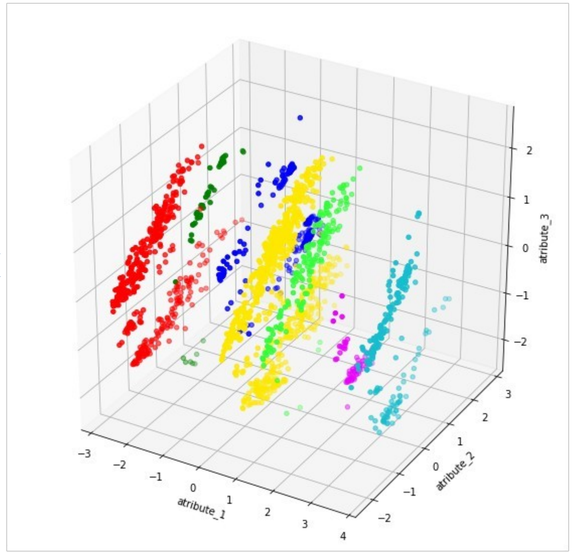


Поиск ортогональных проекций с наибольшим рассеянием

Было



Стало



https://ru.wikipedia.org/wiki/Метод_минимальных_компонент

Логистическая регрессия.

Задача логистической регрессии – определить вероятность принадлежности к классу.

Построена на основе линейной функции.

$$h(x) = \theta^T x$$

К линейной функции применяется функция активации:

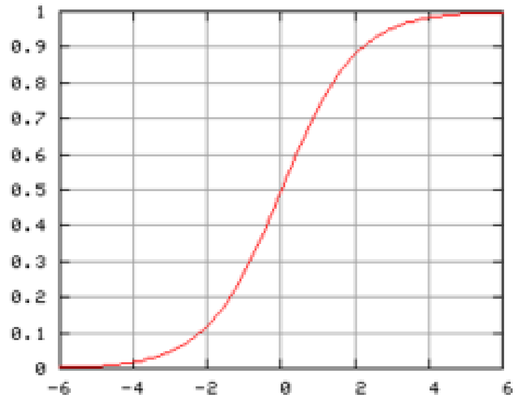
$$h(x) = \sigma(\theta^T x)$$

Функция активации:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

https://towardsdatascience.com/building-a-logistic-regression-in-python-301d27367c24

Сигмоида.



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Производная сигмоиды:

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$$

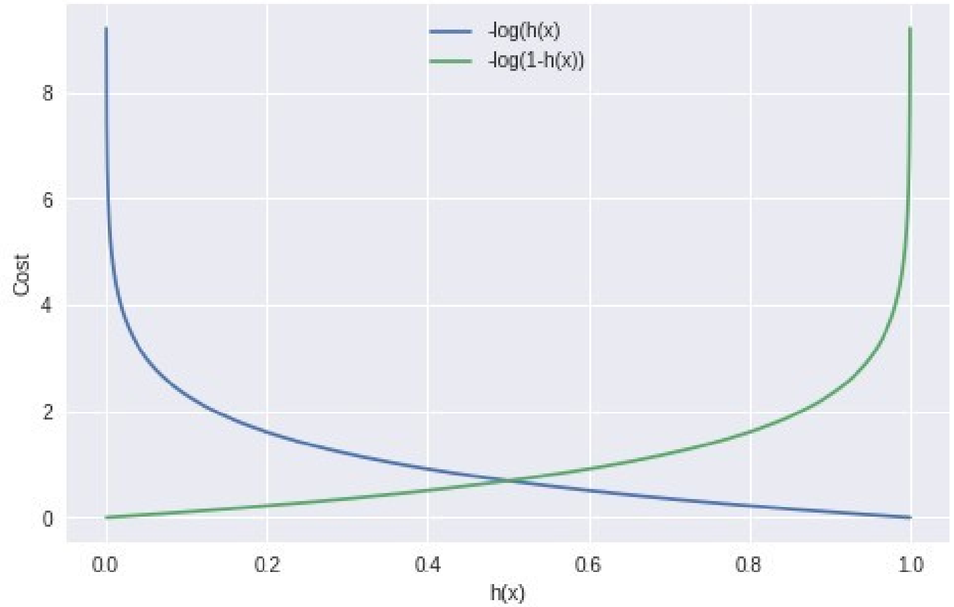
https://ru.wikipedia.org/wiki/Сигмоида

Функция ошибки в логистической регрессии.

Модель ищет параметры, которые минимизируют функцию ошибки:

$$cost = \begin{cases} -log(h(x)), & if \quad y = 1 \\ -log(1 - h(x)), & if \quad y = 0 \end{cases}$$

Чем выше вероятность определения класса 1 при верном классе 0, тем выше стоимость ошибки.



https://towardsdatascience.com/building-a-logistic-regression-in-python-301d27367c24

Функция ошибки в логистической регрессии.

Общий вид функции ошибки для модели:

$$cost(h(x), y) = -y \cdot log(h(x)) - (1 - y)log(1 - h(x))$$

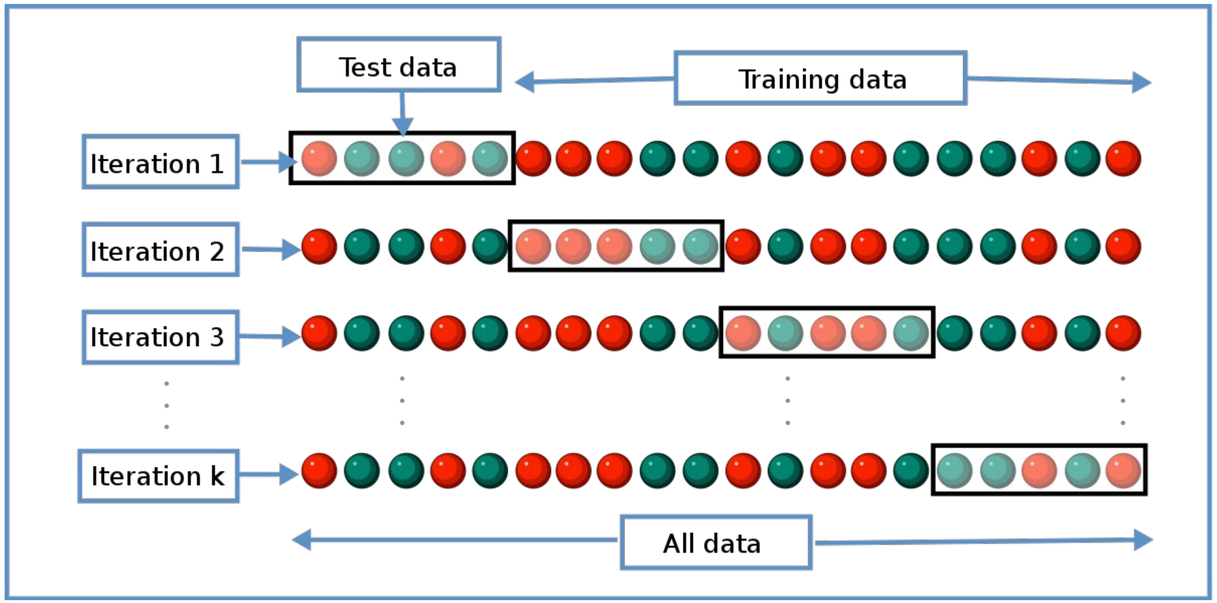
Ошибка для всех данных датасета:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i log(h(x^i)) + (1 - y^i)log(1 - h(x^i))\right]$$

Где m – количество элеметов.

https://towardsdatascience.com/building-a-logistic-regression-in-python-301d27367c24

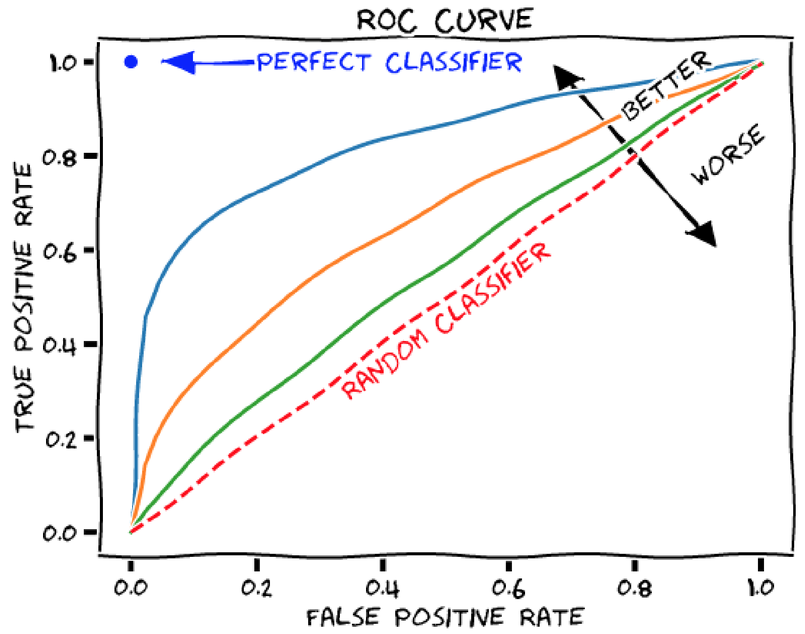
Кросс-валидация



Оцениваем модель на нескольких тестовых данных

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

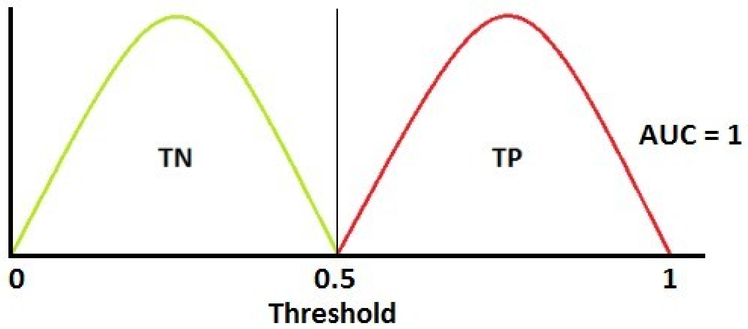
Метрики классификации: ROC-кривая

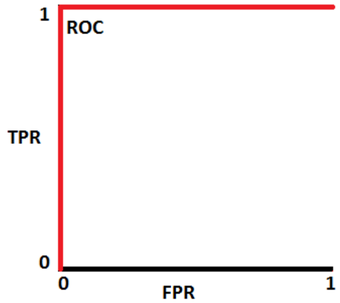


Позволяет определить порог,
при котором мы будем отделять один класс от другого

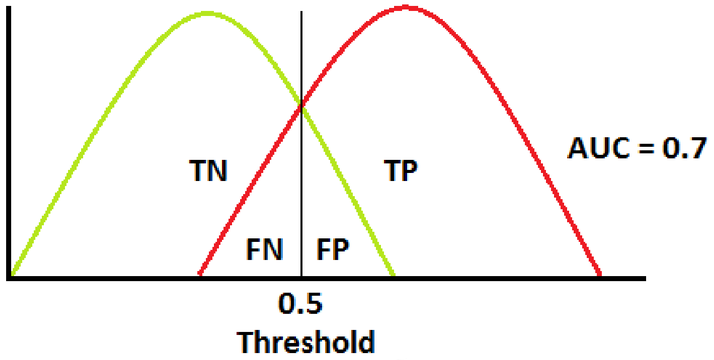
https://analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-to-find-the-perfect-probability-threshold/

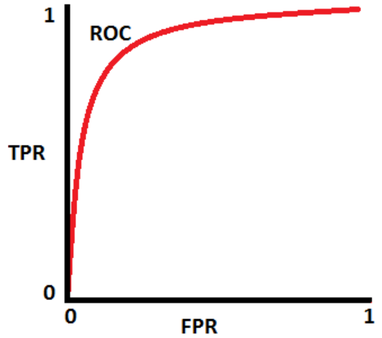
ROC-кривая





Идеальная модель — порог 50%





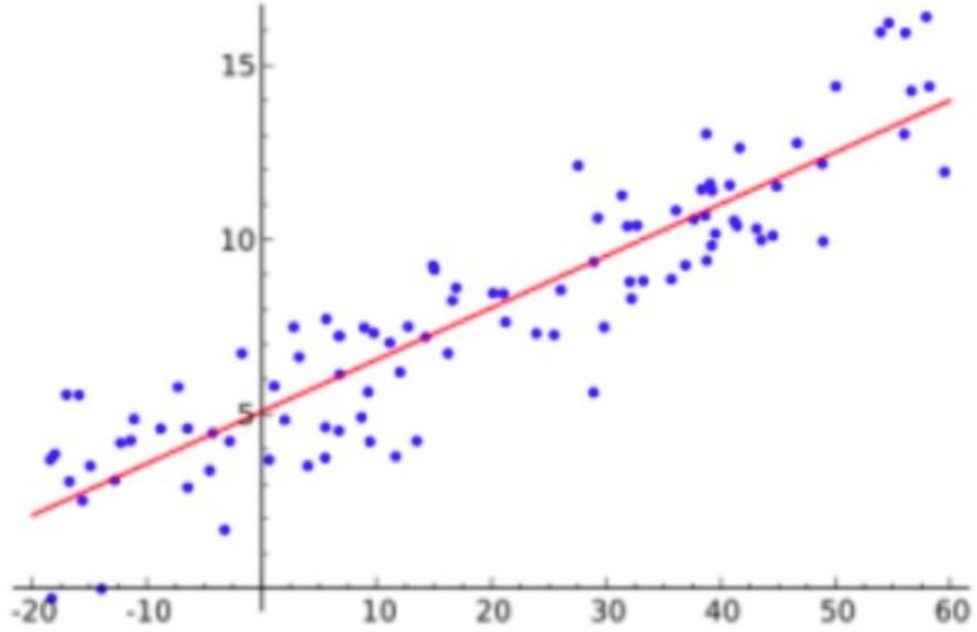
Модель с некотороыми ошибками — порог выбирается в зависимости от допускаемых ошибок

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

НЕТОЛОГИЯ

Регрессия

Отличается тем, что допустимым ответом является действительное число или числовой вектор.



https://github.com/miptmlschool/

Вопросы?

Контакты спикера:
yustiks@gmail.com