НЕТОЛОГИЯ

Юстина Иванова

Программист, data scientist

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

Спикер

НЕТОЛОГИЯ

Юстина Иванова,

- PhD в Унивеситете Больцано
- Data scientist по компьютерному зрению в компании ОЦРВ,
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence В University of Southampton

Классическое Обучение

ОЛОГИЯ

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

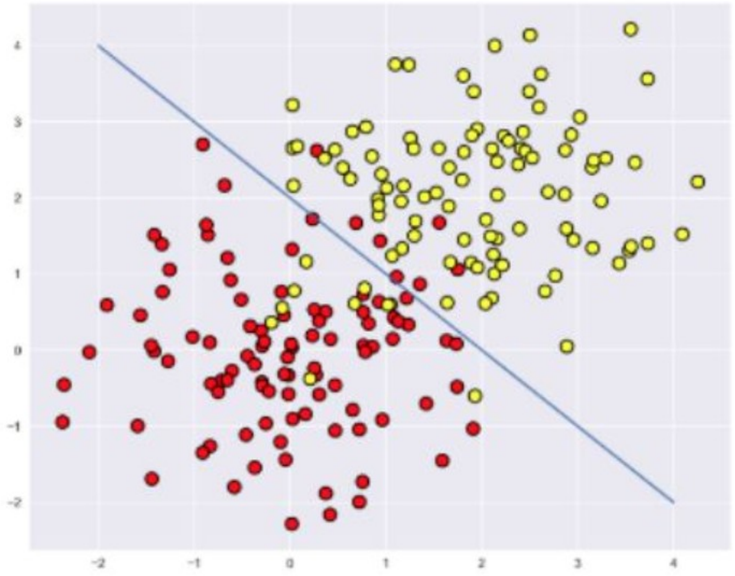
«Собери из вещей лучшие наряды»

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

НЕТОЛОГИЯ

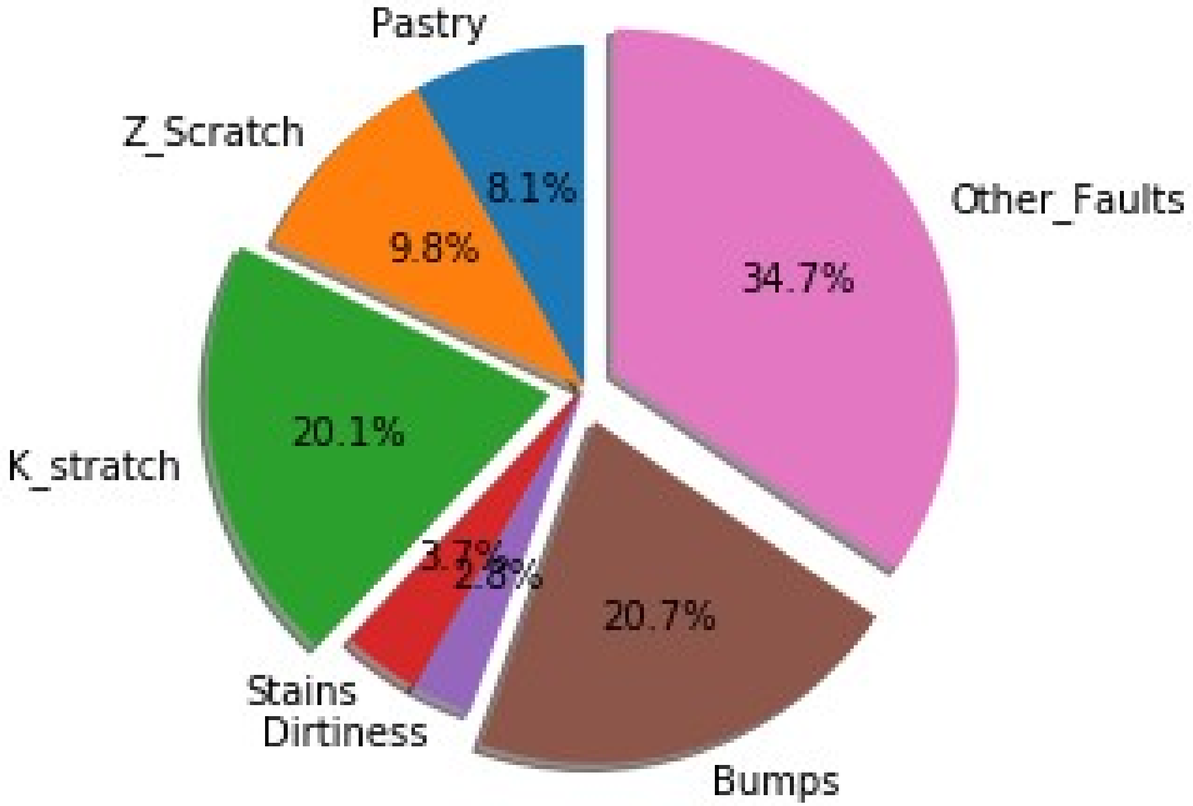
Классификация

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.

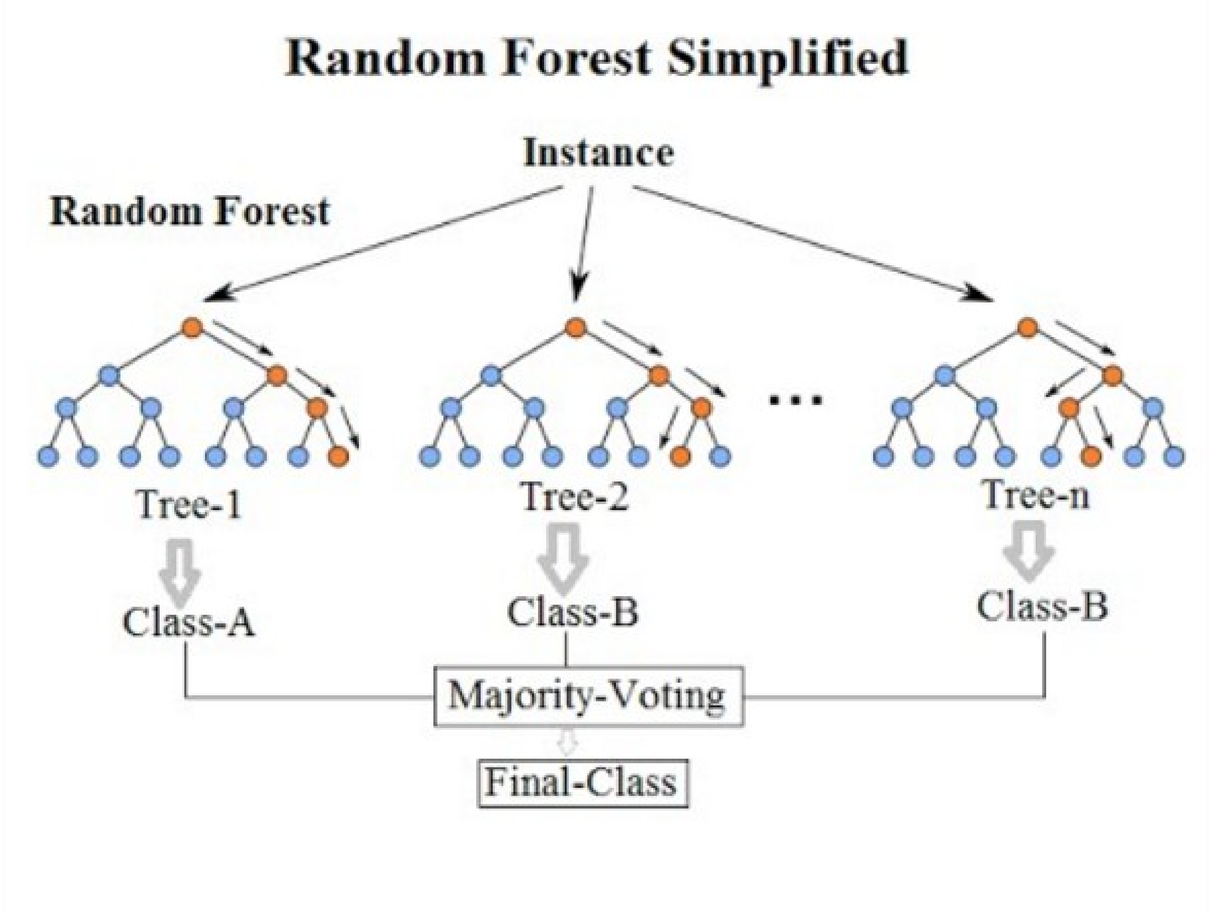


https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

Деревья решений.

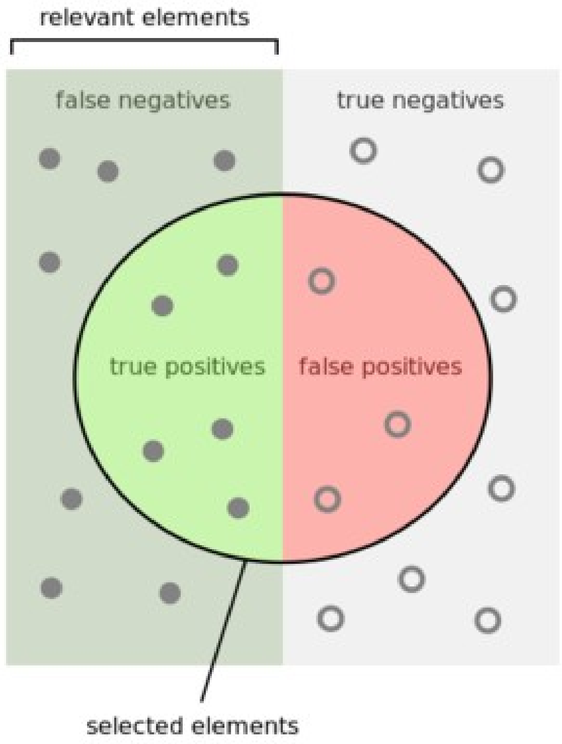
https://vas3k.ru/blog/machine_learning/

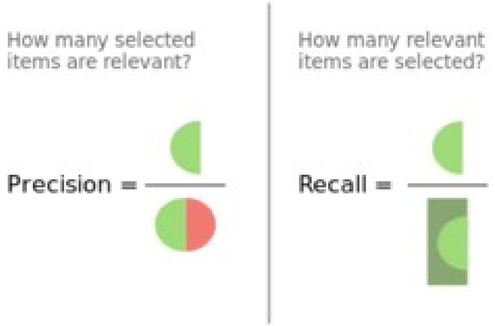
Случайный лес.



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

Метрики классификации





Precision

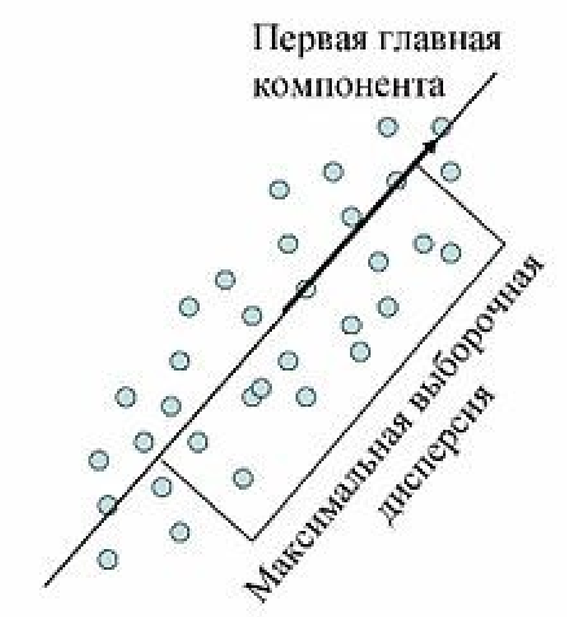
Recall

F1-мера

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

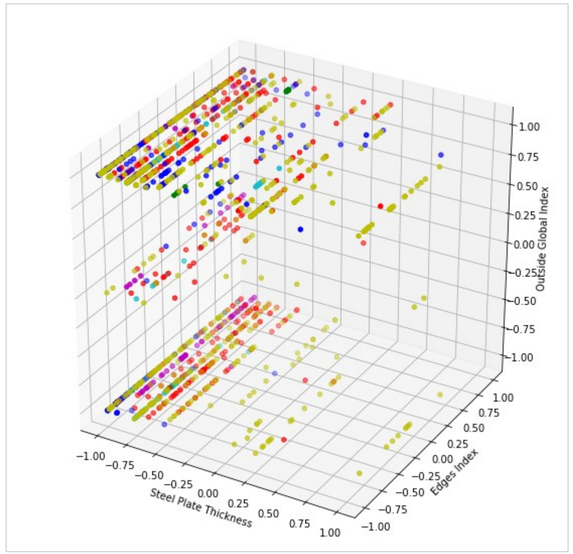
https://habr.com/ru/company/ods/blog/328372/

Принцип минимальных компонент.

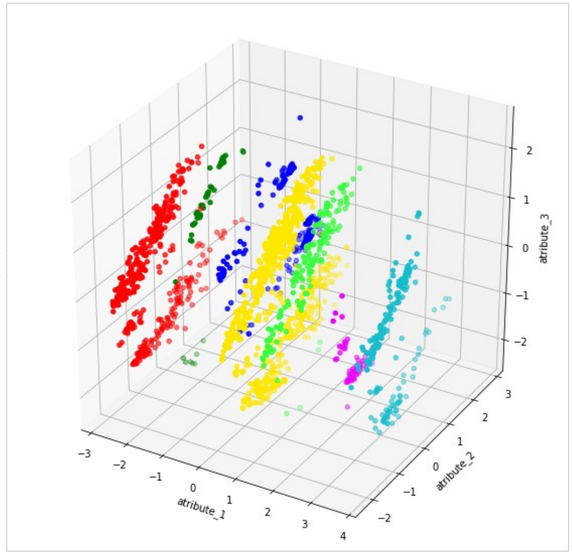


Поиск ортогональных проекций с наибольшим рассеянием

Было

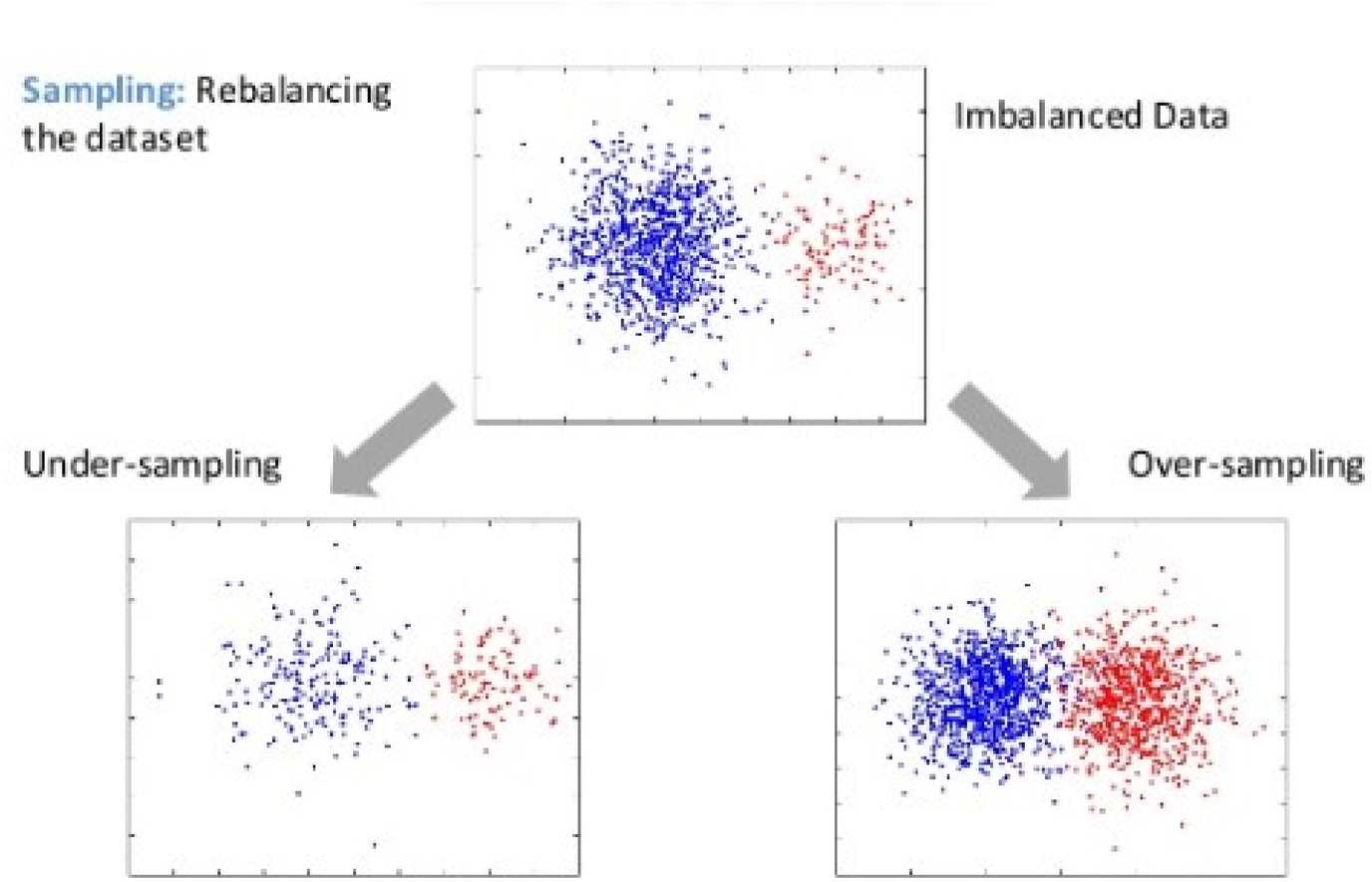


Стало



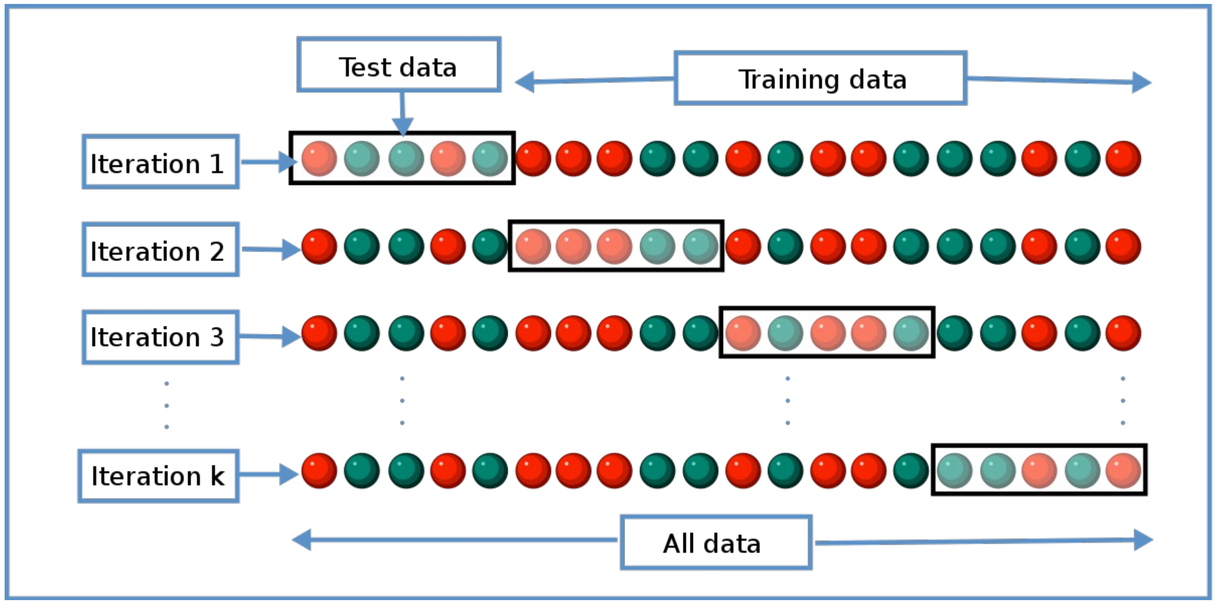
https://ru.wikipedia.org/wiki/Метод_минимальных_компонент

Проблема несбалансированности классов.



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

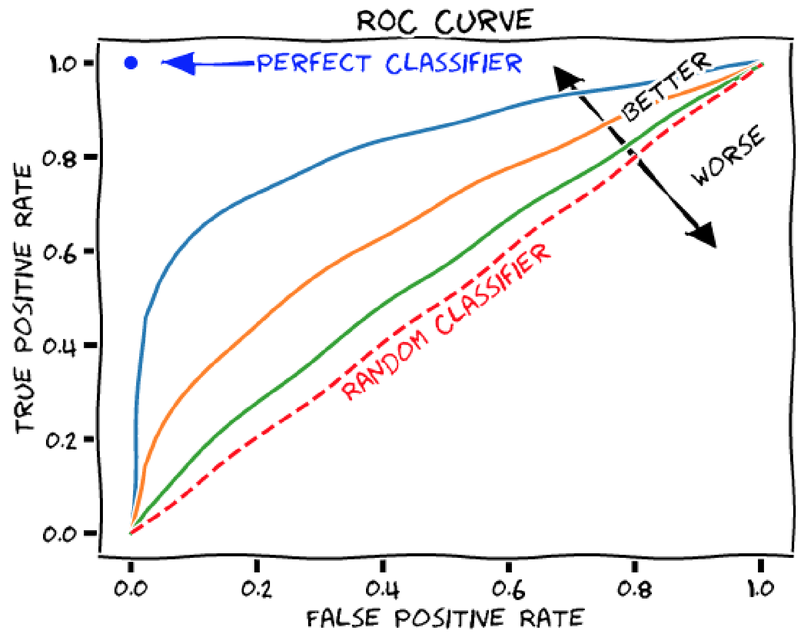
Метрики классификации: кросс-валидация



Оцениваем модель на нескольких тестовых данных

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

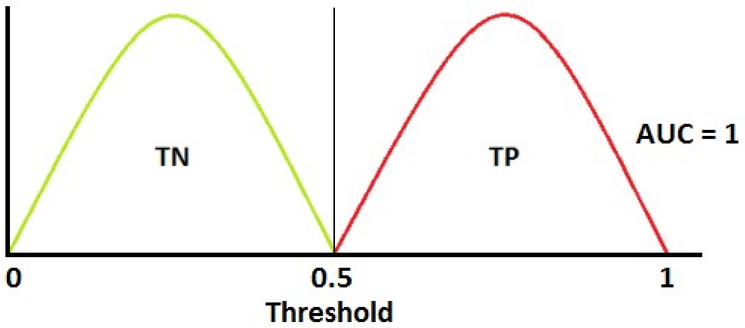
Метрики классификации: ROC-кривая

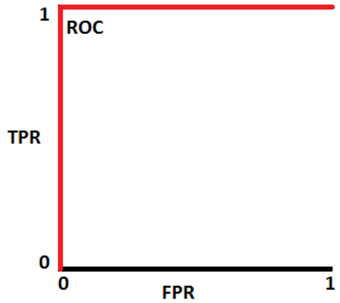


Позволяет определить порог,
при котором мы будем отделять один класс от другого

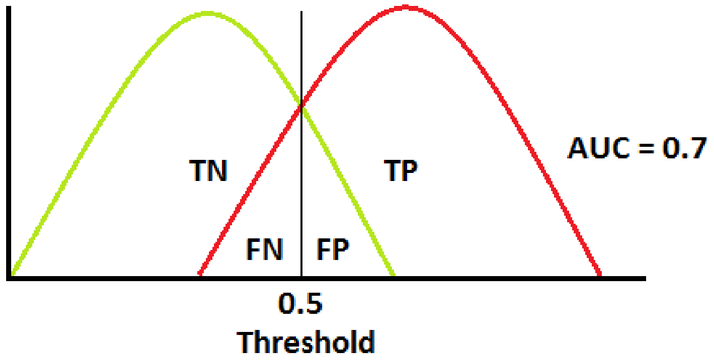
https://analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-to-find-the-perfect-probability-threshold/

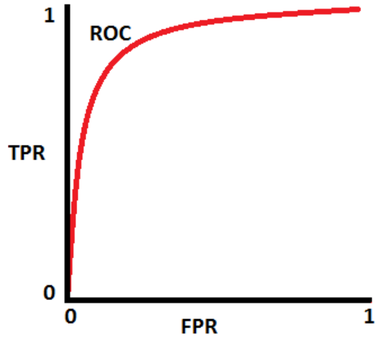
ROC-кривая





Идеальная модель — порог 50%





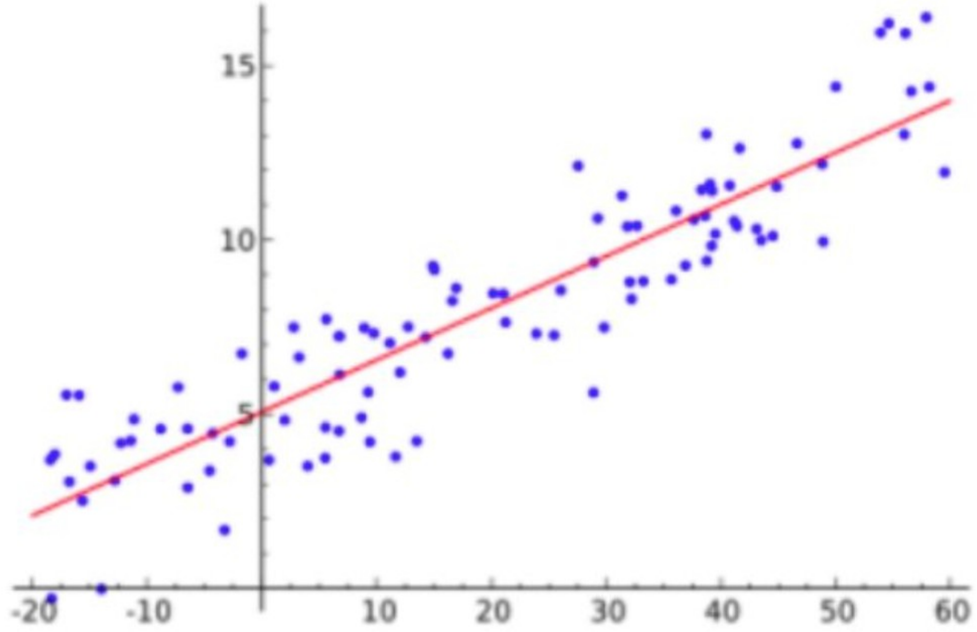
Модель с некотороыми ошибками — порог выбирается в зависимости от допускаемых ошибок

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

НЕТОЛОГИЯ

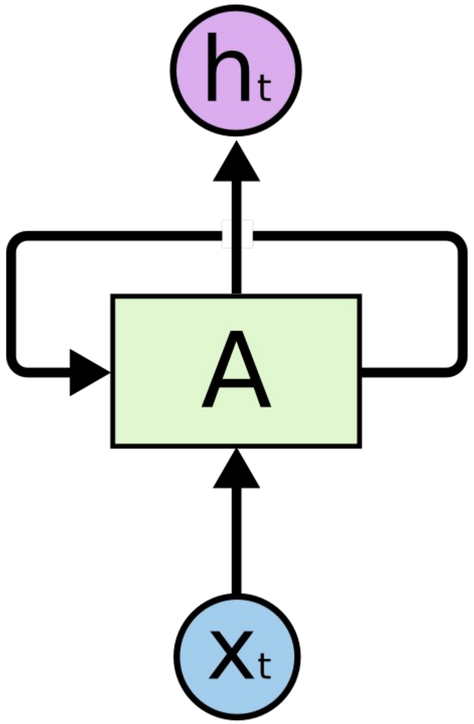
Регрессия

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



https://github.com/miptmlschool/

Модель LSTM.

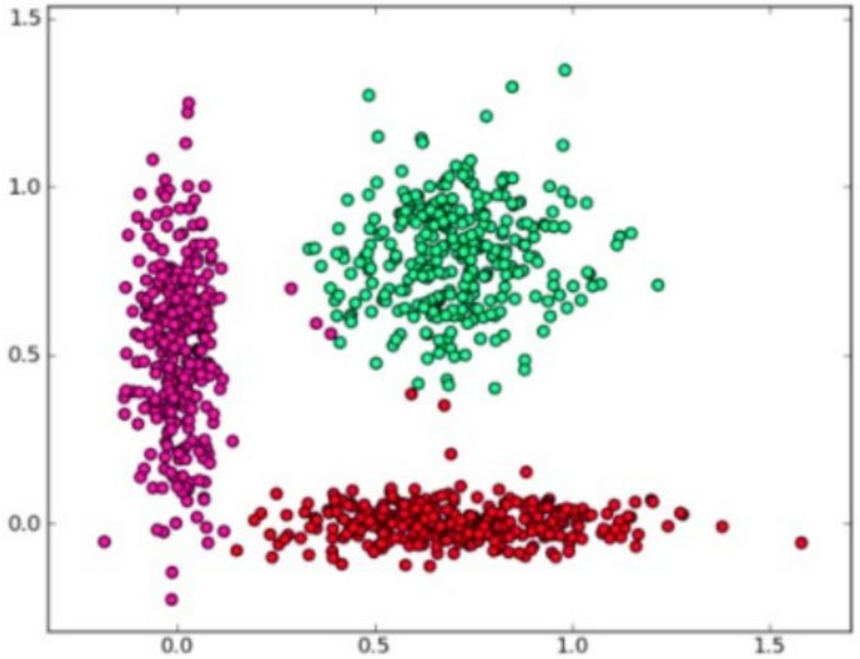


Рекуррентные нейронные сети содержат обратные связи.

https://habr.com/ru/company/wunderfund/blog/331310/

Кластеризация

Заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.



https://github.com/miptmlschool/

Вопросы?

Контакты спикера:
yustiks@gmail.com