НЕТОЛОГИЯ

Юстина Иванова

Программист, data scientist

Статистика в python. Кейс-стади №1.
Датасеты: faulty steel plates,
Titanic, Boston houses.

Спикер

НЕТОЛОГИЯ

Юстина Иванова,
Data scientist по
Компьютерному зрению
в компании ОЦРВ,
Выпускница МГТУ им. Баумана
Магистр по Artificial Intelligence
В University of Southampton

Классификационный анализ

НЕТОЛОГИЯ

Среди самых популярных задач Т в машинном обучении:

классификация – отнесение объекта к одной из категорий на основании его признаков
регрессия – прогнозирование количественного признака объекта на основании прочих его признаков
кластеризация – разбиение множества объектов на группы на основании признаков этих объектов так, чтобы внутри групп объекты были похожи между собой, а вне одной группы – менее похожи
детекция аномалий – поиск объектов, "сильно непохожих" на все остальные в выборке либо на какую-то группу объектов
и много других, более специфичных.
Хороший обзор дан в главе "Machine Learning basics" книги "Deep Learning" (Ian Goodfellow, Yoshua Bengio, Aaron Courville, 2016)

https://habr.com/ru/company/ods/blog/322534/

НЕТОЛОГИЯ

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:

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

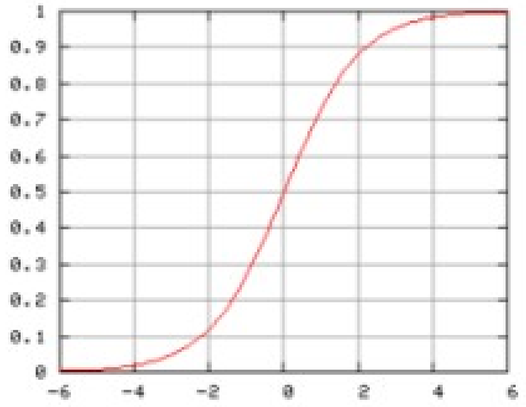
Функция активации:

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

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

НЕТОЛОГИЯ

Сигмоида.



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

Производная сигмоиды:

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

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

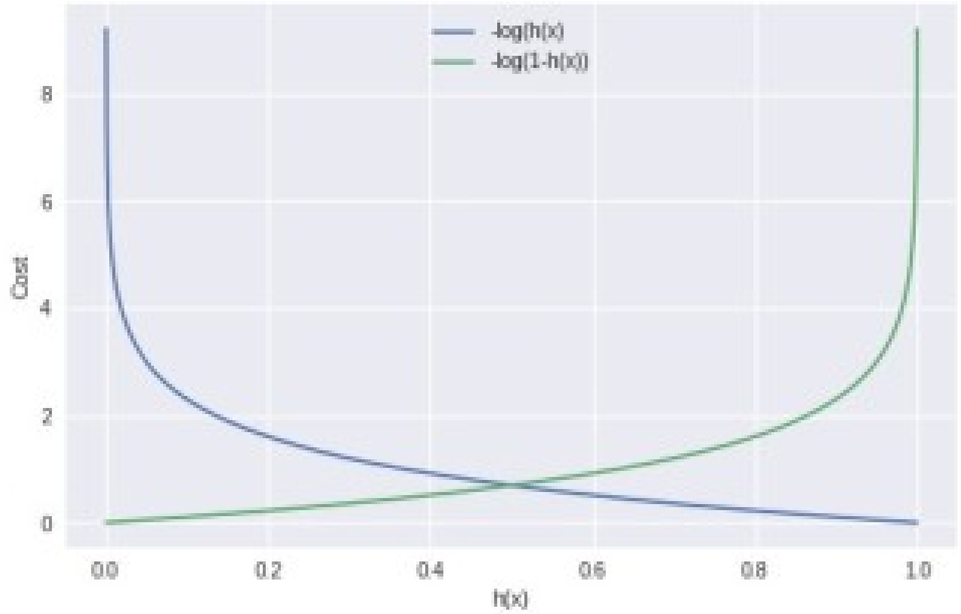
НЕТОЛОГИЯ

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

НЕТОЛОГИЯ

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

НЕТОЛОГИЯ

Градиентный спуск.

Будем искать минимум функции относительно параметров:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h(x^i) - y^i) x_j^i$$

Для поиска минимума используется градиентный спуск.

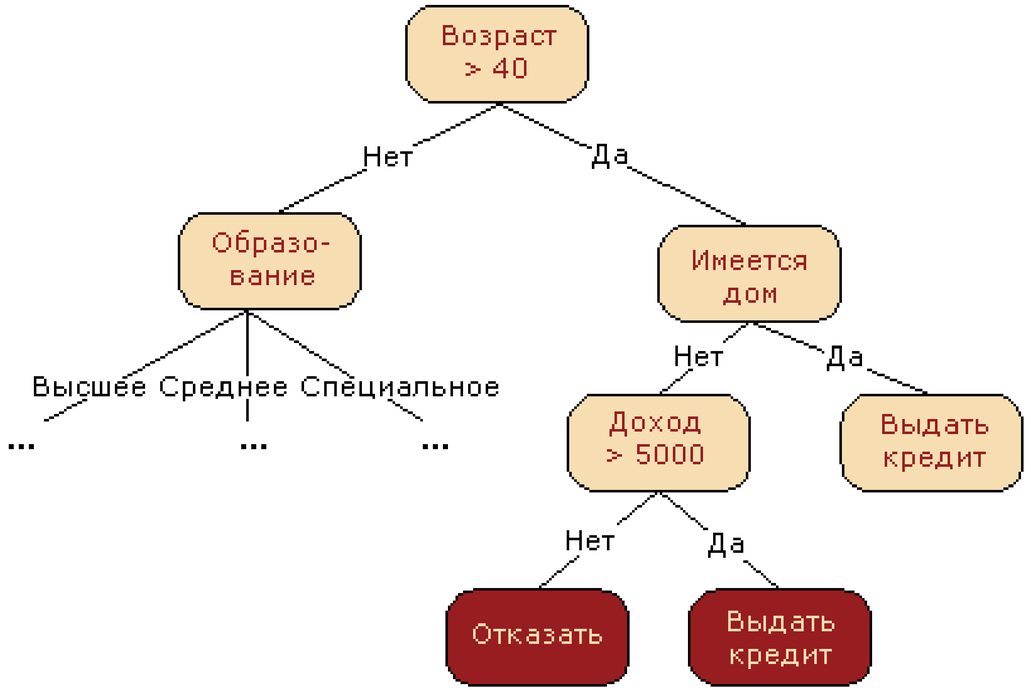
Это метод нахождения локального минимума (максимума) функции с помощью движения вдоль градиента.

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

Деревья решений.

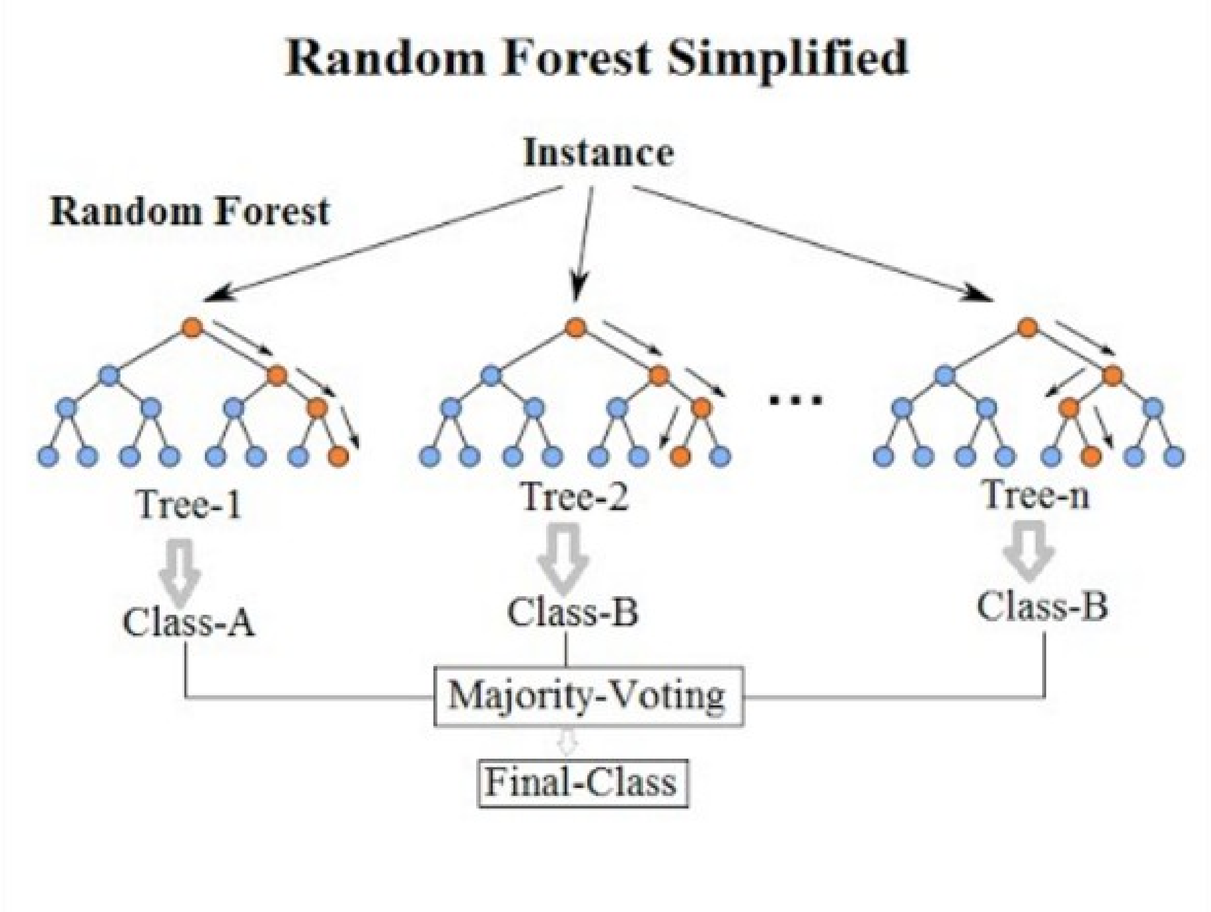
https://ru.wikipedia.org/wiki/Дерево_решений

Деревья решений.



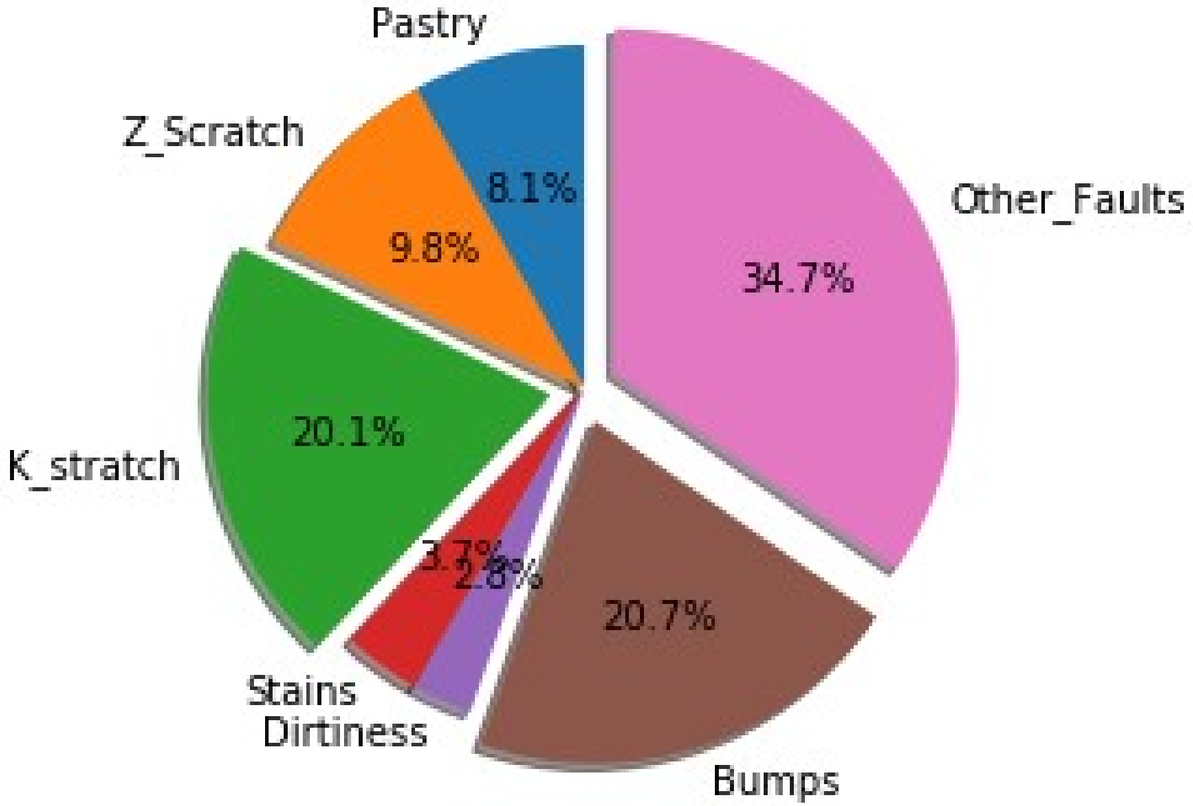
https://habr.com/ru/company/ods/blog/322534/

Случайный лес.



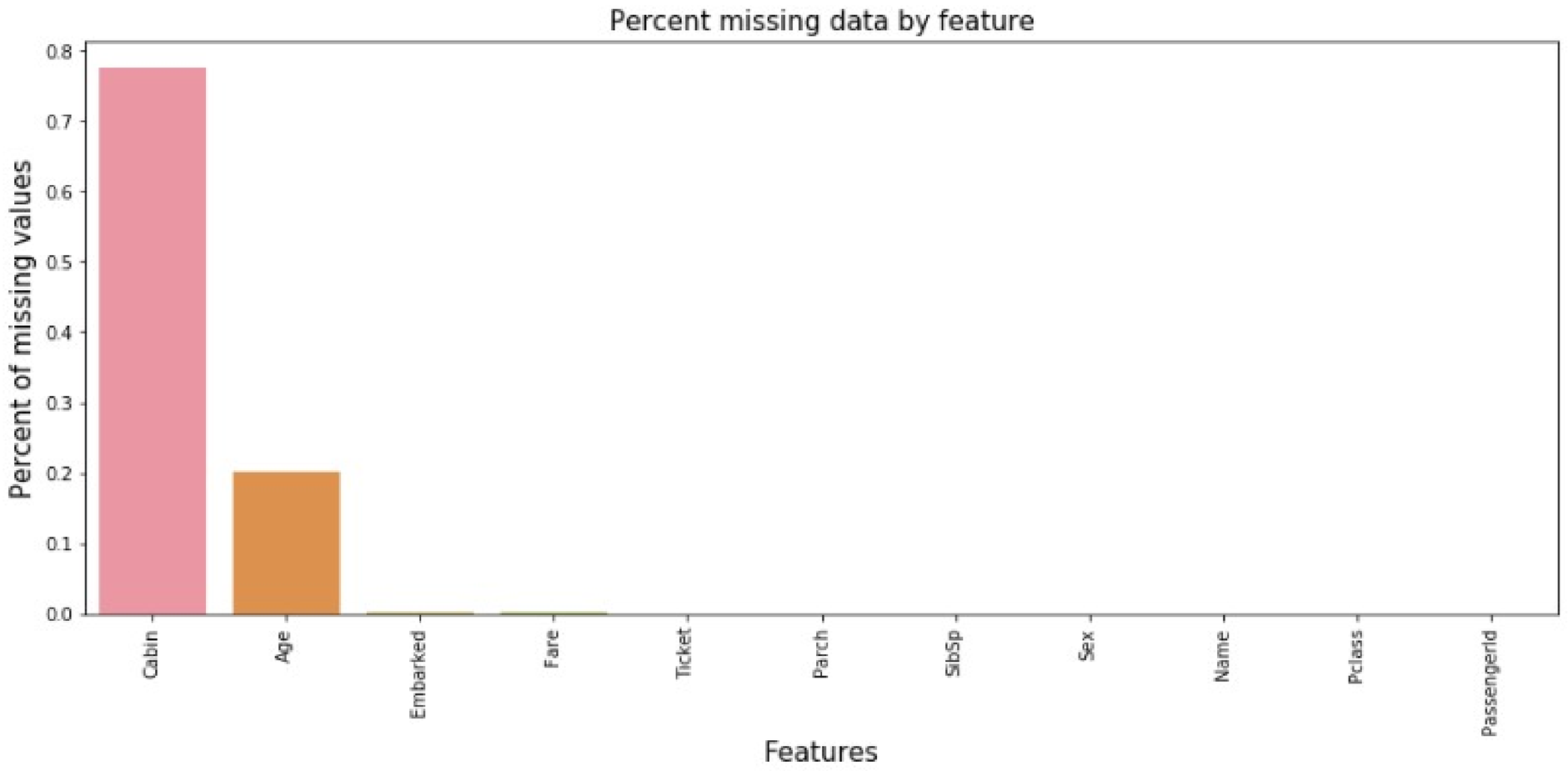
https://ru.wikipedia.org/wiki/Random_forest

Проблема несбалансированности классов.



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

Проблема нехватки информации для элементов.



https://towardsdatascience.com/handling-missing-values-in-machine-learning-part-1-dda69d4f88ca

Методы заполнения недостающих данных.

- Удаление элементов
- Заполнение данными предыдущего или последующего элемента
- Заполнение некой константой (выходящей за пределы интервала значений)
- Заполнение средним значением или модой
- Создать атрибут: отсутствующее значение

https://towardsdatascience.com/handling-missing-values-in-machine-learning-part-1-dda69d4f88ca

Вопросы?

Контакты спикера:
yustiks@gmail.com