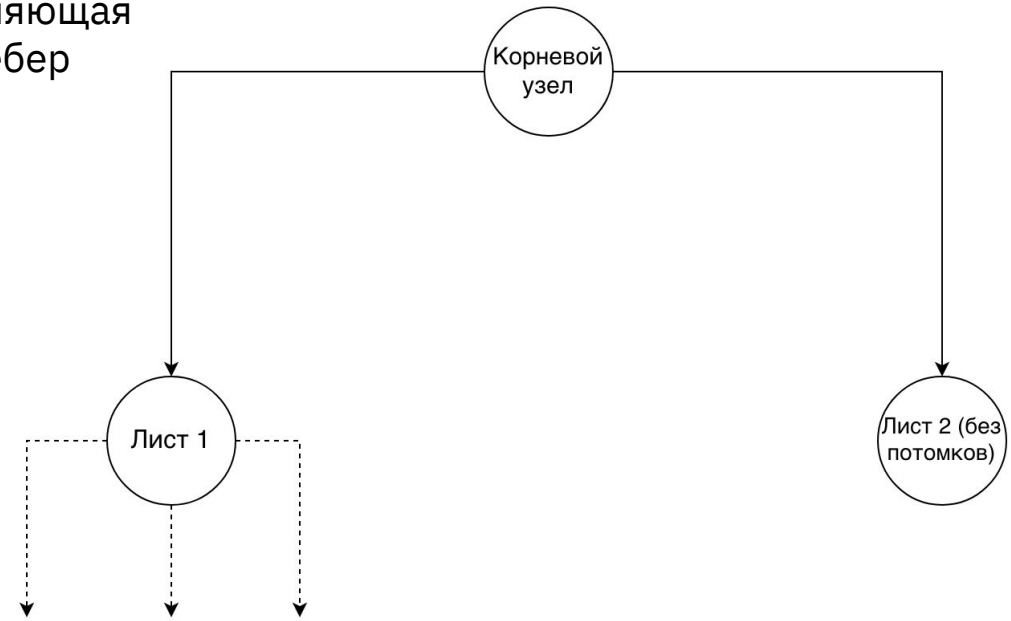
Решающие деревья. Предобработка данных

Что будет на уроке?

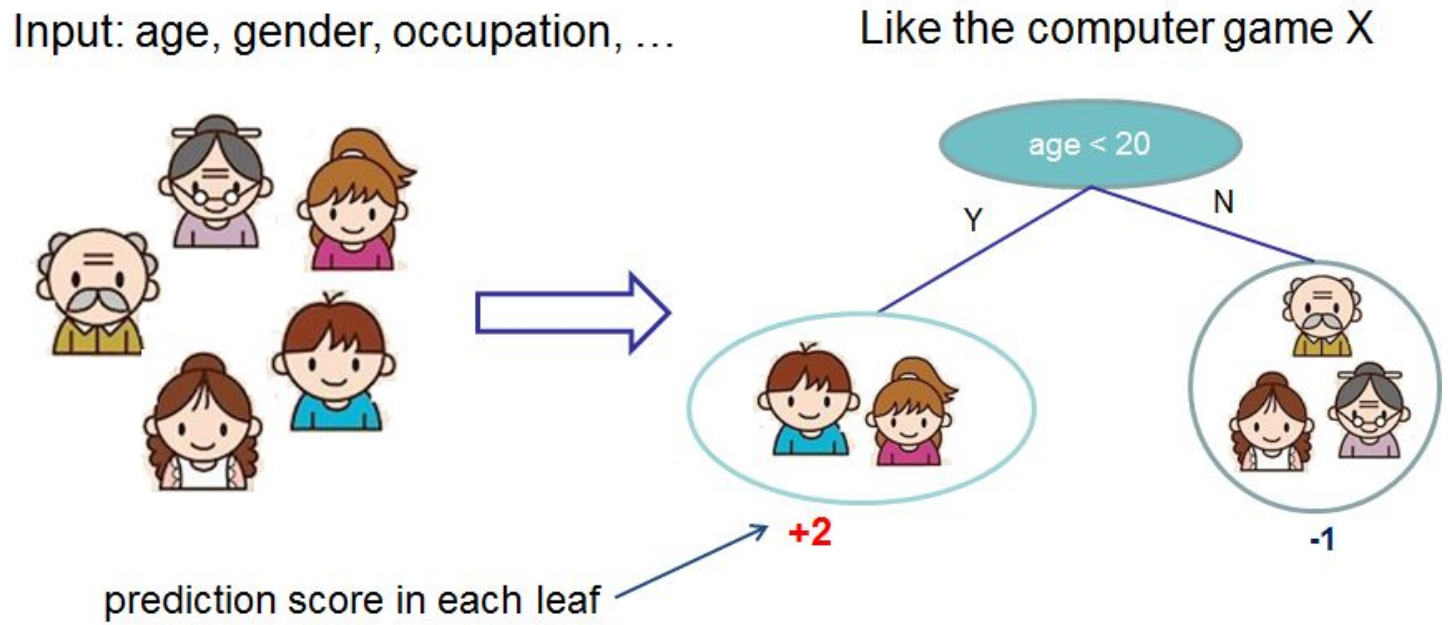
1. Что такое решающее дерево.
2. Алгоритмы, основанные на деревьях.
3. Предобработка данных.
4. Что делать при несбалансированных классах.

Что такое дерево?

Дерево — структура данных, представляющая собой набор вершин (корень, лист) и рёбер (ветвей).



Что такое решающее дерево?



src: https://xgboost.readthedocs.io/en/latest/tutorials/model.html

Плюсы и минусы алгоритмов на деревьях

Плюсы

- Легко интерпретируются.
- Подходят для задачи регрессии и для задачи классификации.
- Не нуждаются в масштабировании признаков.
- Проверяется статистическими тестами.

Минусы

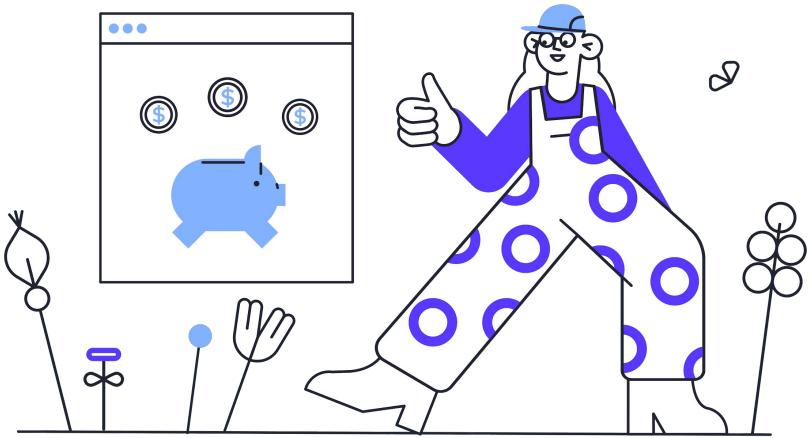
- Склонны к переобучению.
- Плохо работают на несбалансированных данных.

Ансамбли деревьев

Цель ансамблевых методов — объединять предсказания нескольких эстиматоров, построенных с заданным алгоритмом обучения, чтобы улучшить их обобщаемость или устойчивость.

Различают два основных ансамблевых метода:

1. Усредняющие методы (Averaging methods).
2. Методы, основанные на технике бустинга (Boosted methods).



Усредняющие методы

Суть: независимо друг от друга обучаем несколько эстиматоров, затем усредняем их предсказания.

Примеры таких методов:

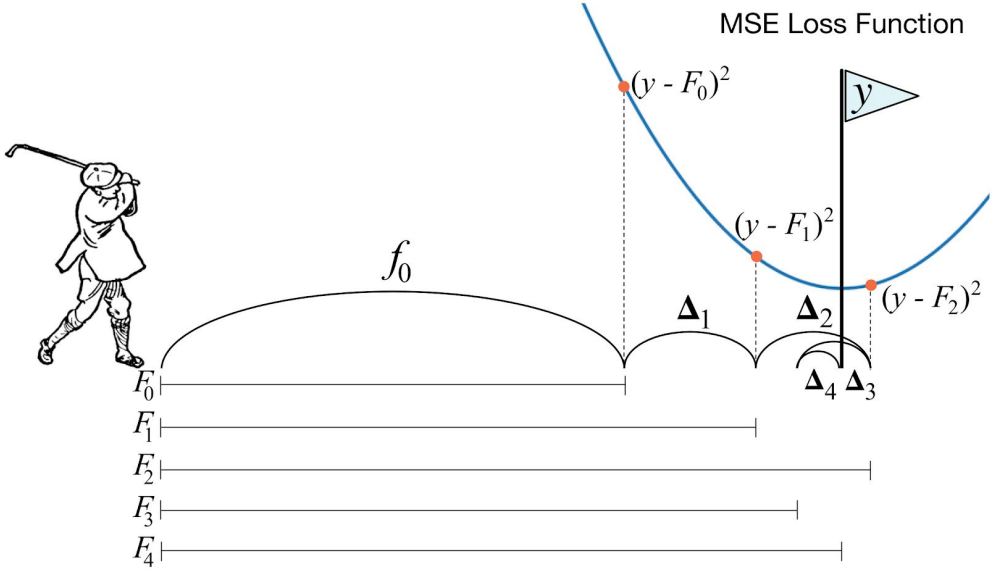
1. Bagging.
2. Случайные леса (Random forests).

Boosted methods

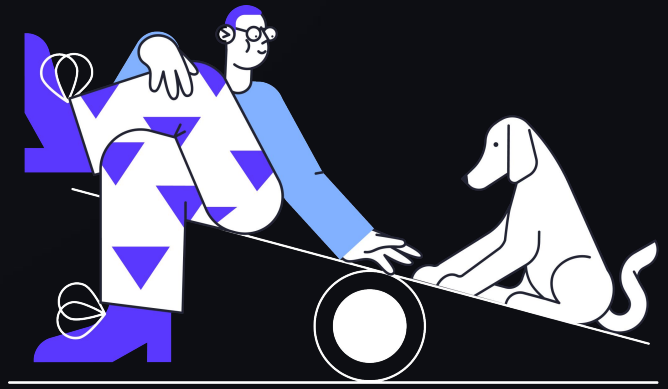
Суть: эстиматоры строятся один за другим, каждый последующий обучается на ошибке предыдущего.

Примеры таких методов:

1. AdaBoost.
2. GradientBoostedTrees.



Предобработка признаков



Практическое задание

1. Изучите методические материалы к занятию.
2. Пройдите тест с выбором варианта ответа.