IT-Academy

Ансамбли моделей. Ч1.

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

Вопросы занятия

1. Классификаторы с голосованием;

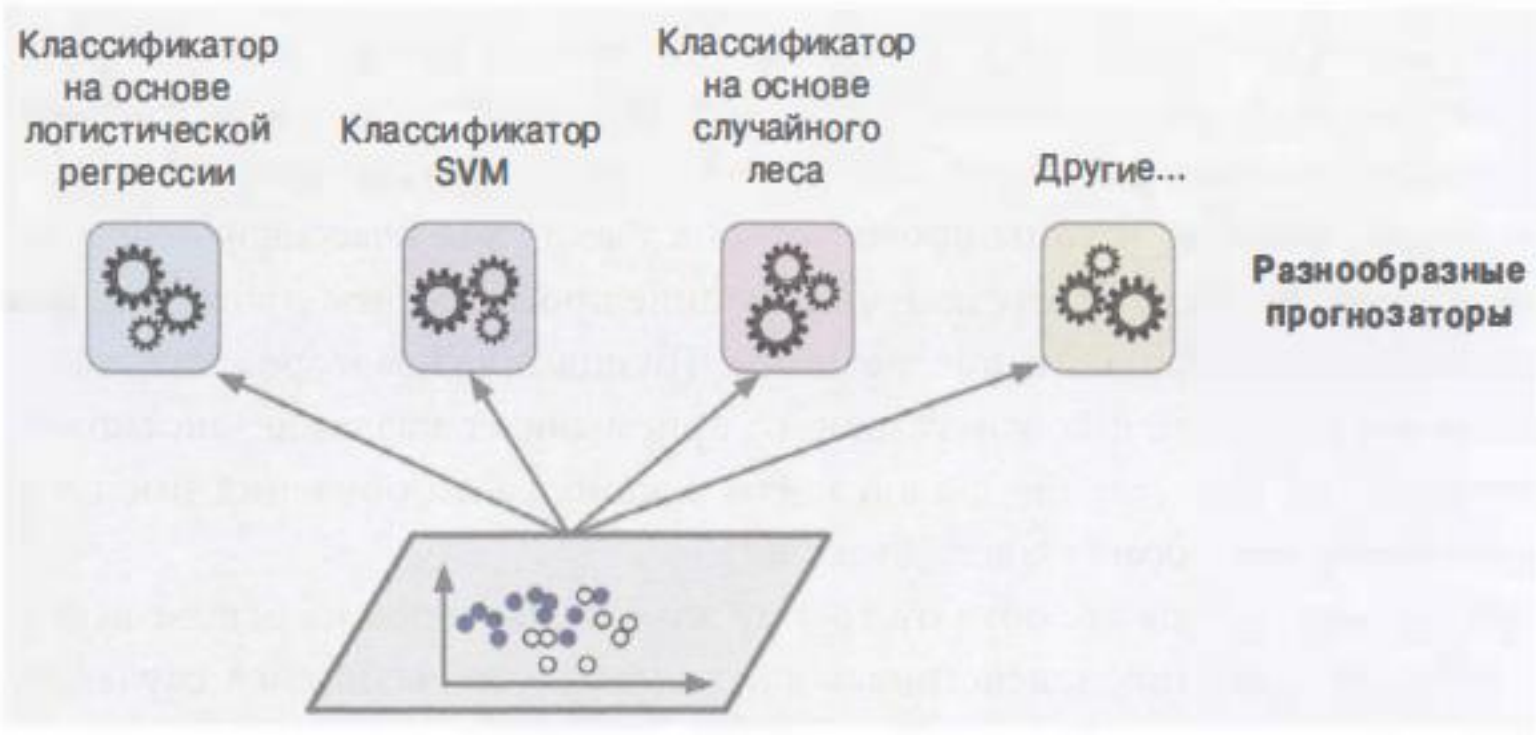
2. Бэггинг и вставка;

3. Случайный лес.

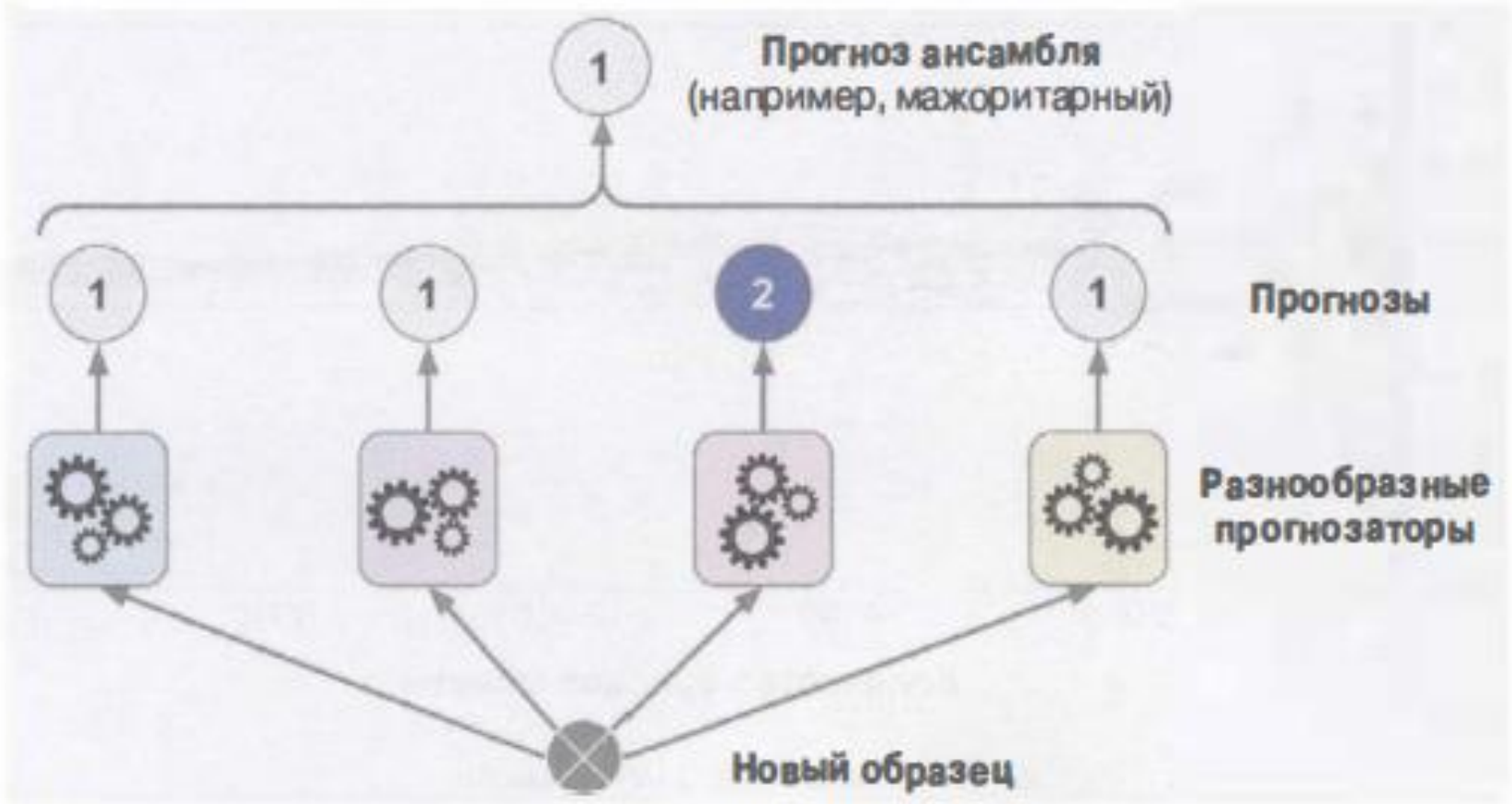
В конце занятия научимся:

- бороться с переобучением при помощи ансамблей;
- объяснять алгоритм случайного леса и использовать его в реальных задачах;
- интерпретировать ML модель, автоматически оценивая важность признаков.

Классификаторы с голосованием



Классификаторы с голосованием



IT-Academy

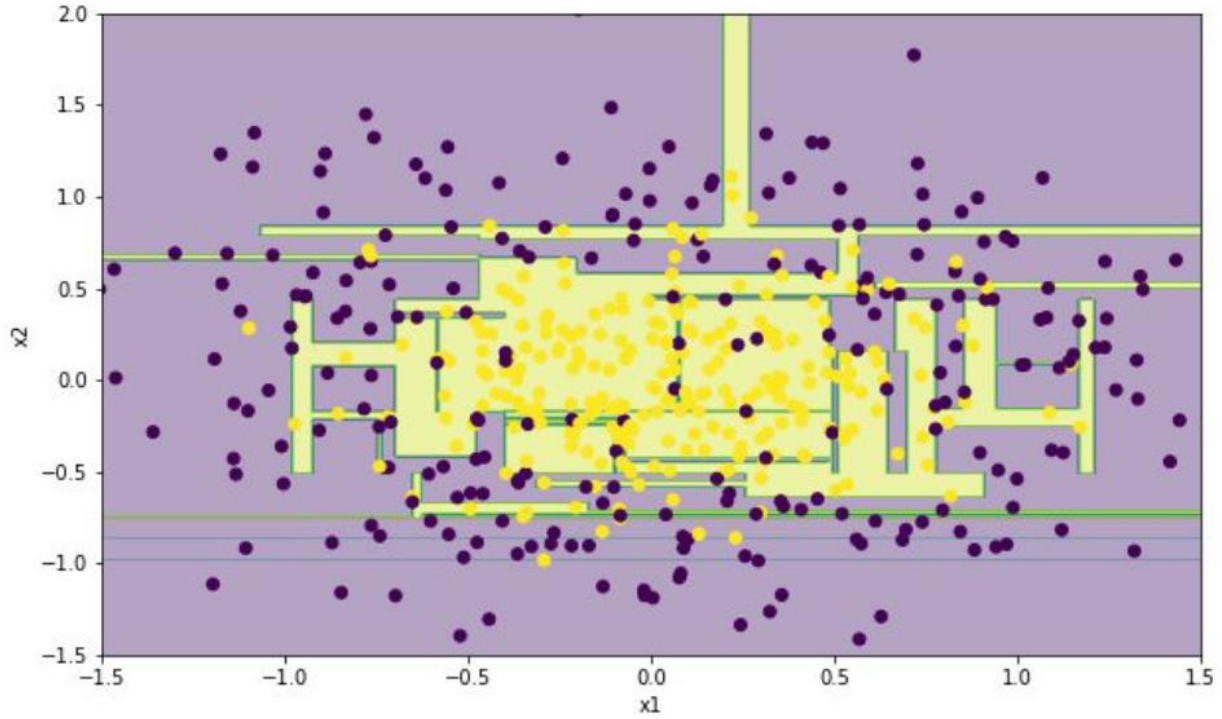
ПРАКТИКА

Voting.ipynb

Проблем с некоторыми моделями

IT-Academy

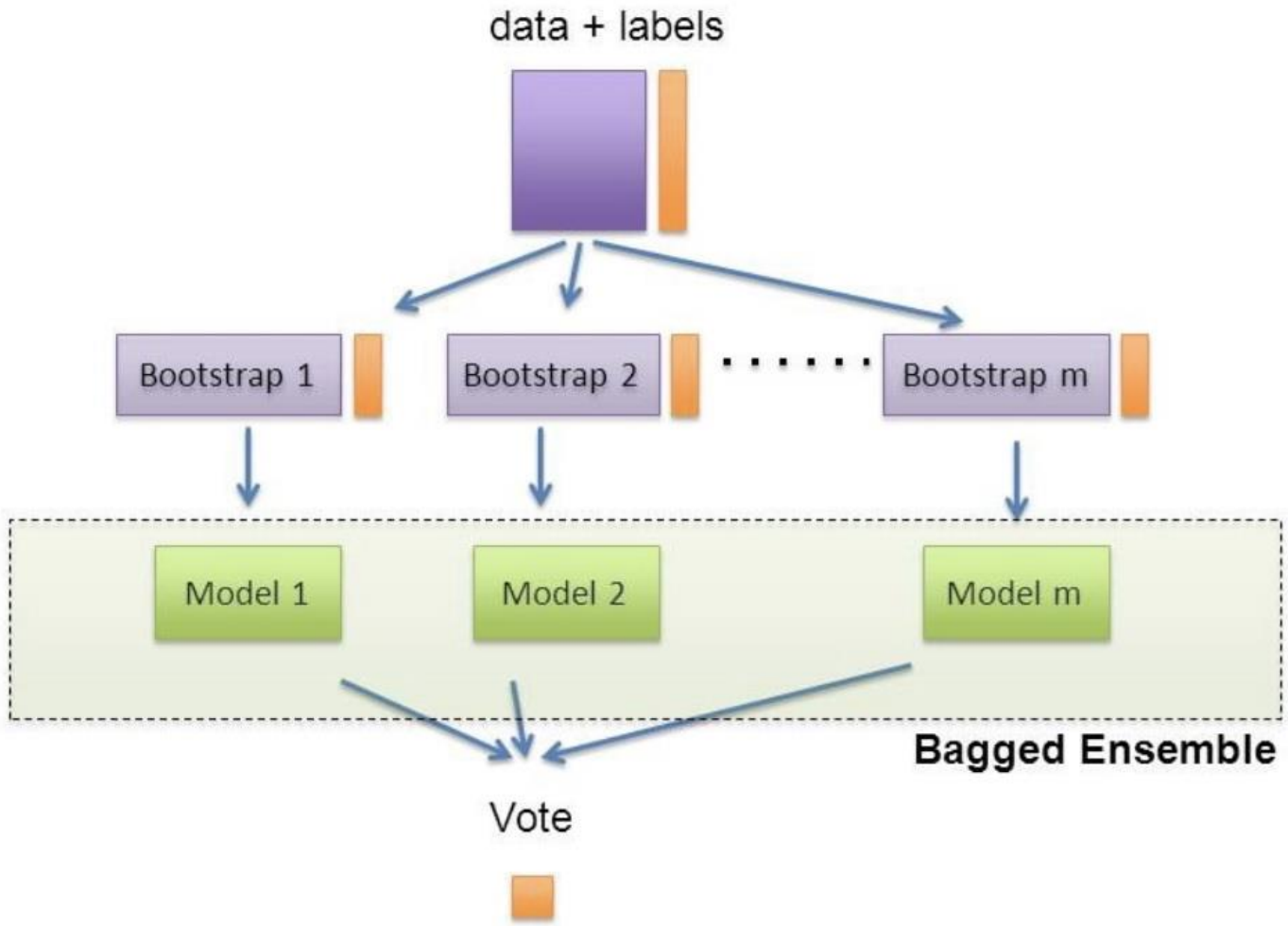
НЕЛИНЕЙНЫЕ МОДЕЛИ ЧАСТО ПЕРЕОБУЧАЮТСЯ



Бэггинг

IT-Academy

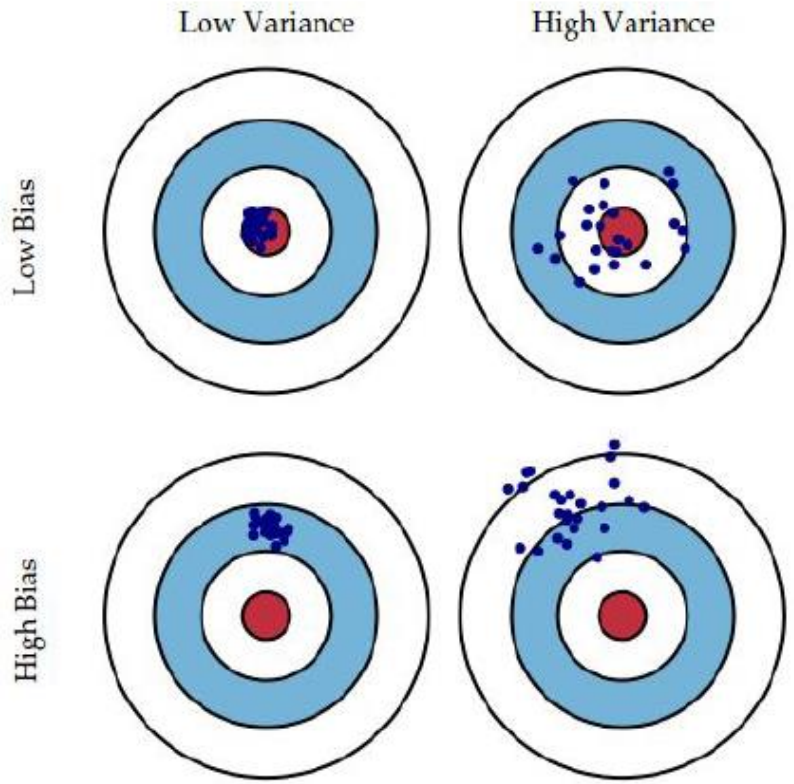
BAGGING =
BOOTSTRAP
AGGREGATION



Бэггинг

IT-Academy

ПОВТОРЕНИЕ: BIAS/VARIANCE



Бэггинг

IT-Academy

АЛГОРИТМ

Дано: выборка ***X*** размера ***N***

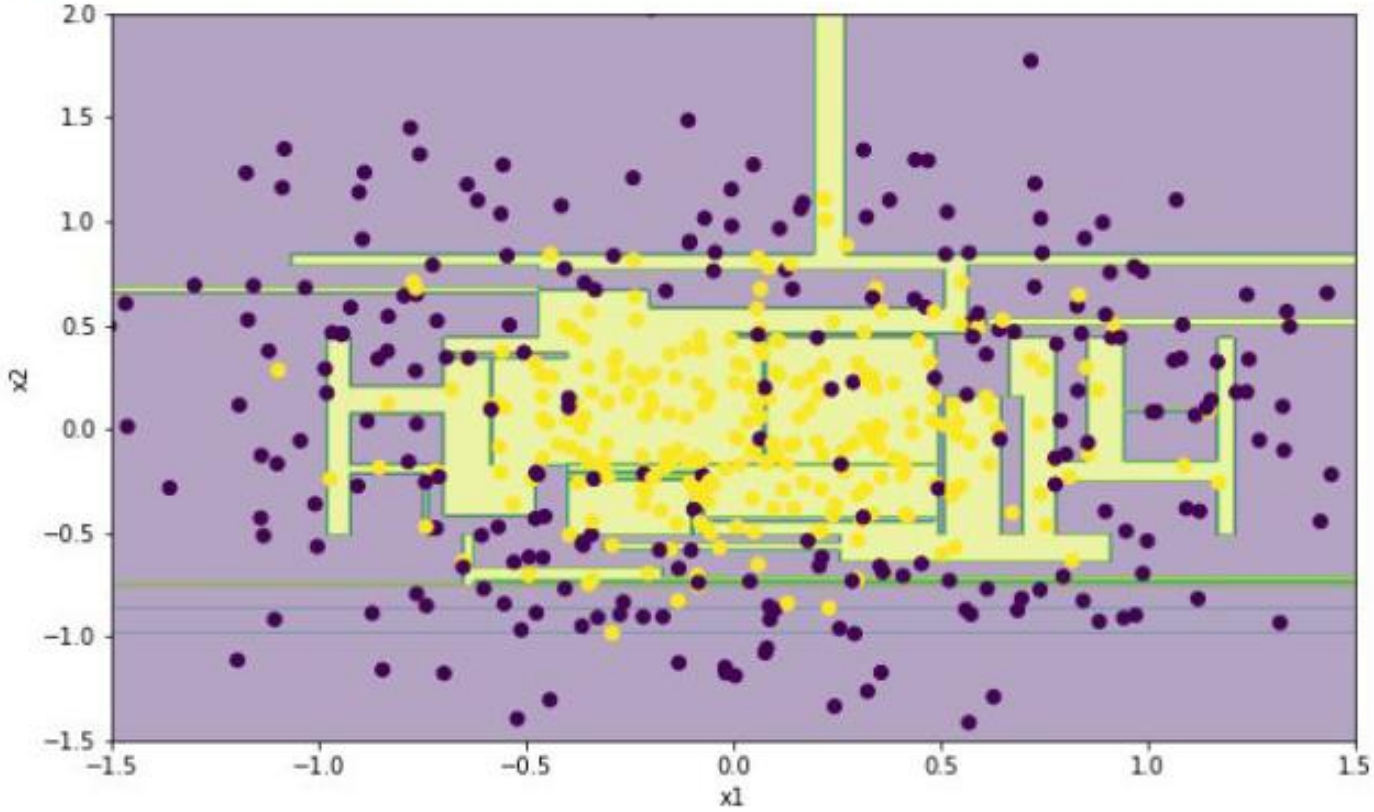
1. Генерируем подвыборку X_i размера ***N*** с возвращением
2. Обучим базовый алгоритм $a_i(x)$ на выборке
3. Повторяем шаги 1-2 ***M*** раз
4. Усредняем (регрессия) или проводим голосование среди ответов $a_i(x)$:

$$a(x) = \frac{1}{M} \sum_{i=1}^{M} a_i(x)$$

Бэггинг

РАЗДЕЛЯЮЩАЯ ПОВЕРХНОСТЬ

Решающее дерево

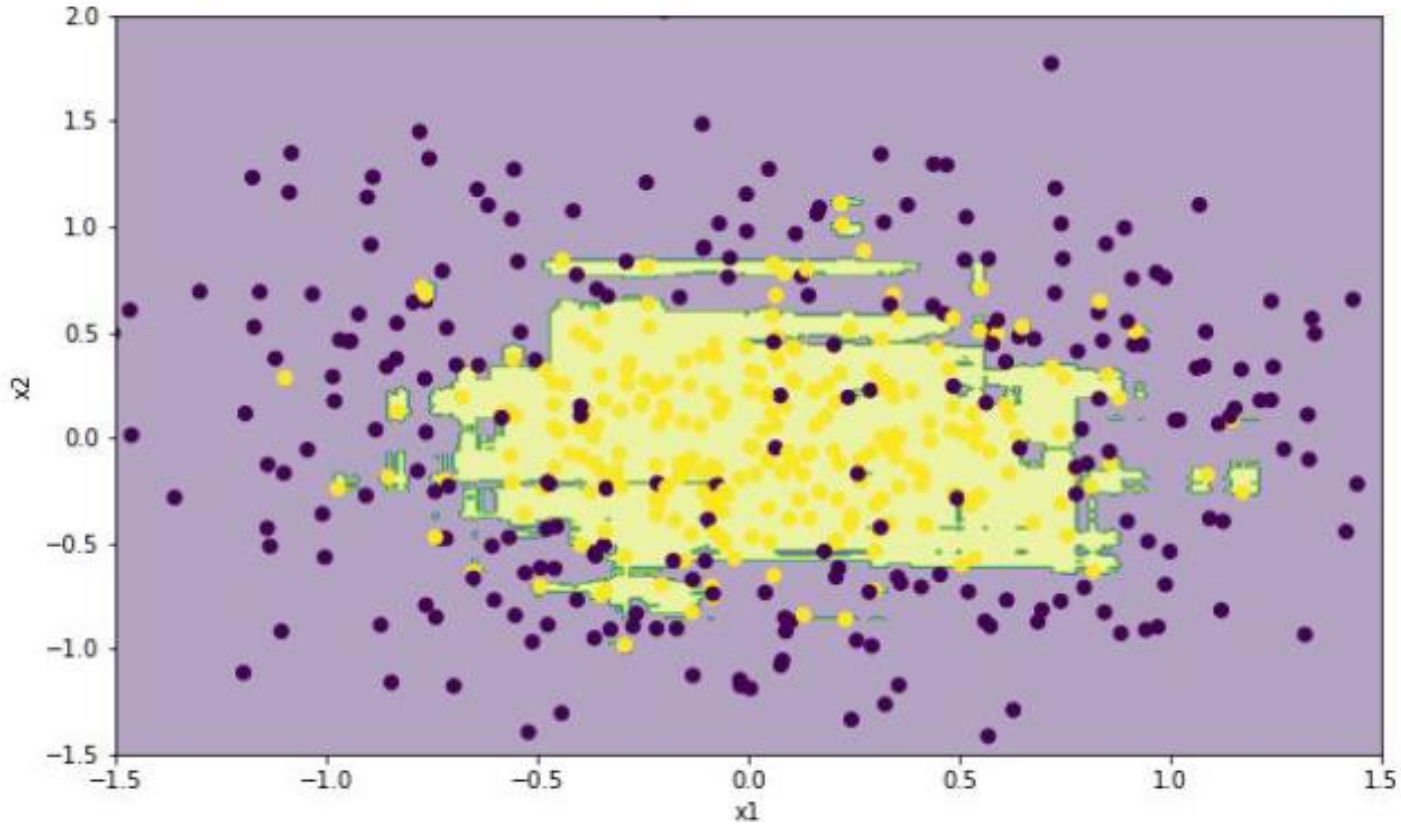


Бэггинг

IT-Academy

РАЗДЕЛЯЮЩАЯ ПОВЕРХНОСТЬ

Бэггинг 300 решающих деревьев

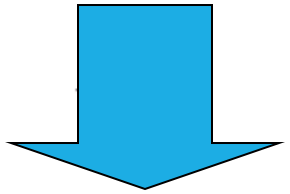


Бэггинг

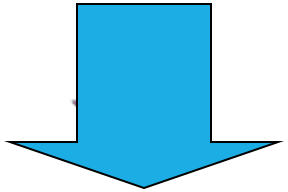
IT-Academy

ВАЖНО

минимальная корреляция ошибок базовых алгоритмов



строим вариативные, неустойчивые модели

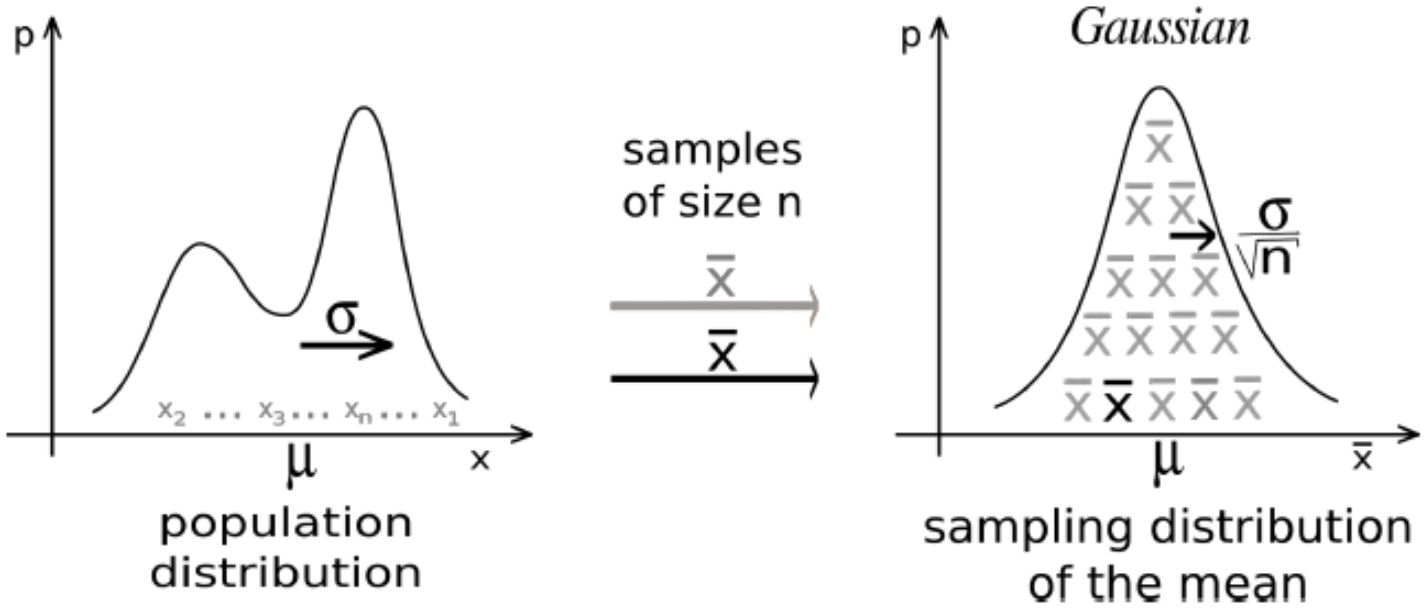


решающие деревья отличный выбор!

Бэггинг

ПОЧЕМУ УМЕНЬШАЕТСЯ VARIANCE?

Центральная предельная теорема:



*википедия, центральная предельная теорема

Бэггинг

IT-Academy

ОСОБЕННОСТИ

– уменьшает variance модели;

– заметно улучшает качество нестабильных базовых алгоритмов;

– может увеличить bias и ухудшить качество стабильного алгоритма, т.к. в каждой подвыборке в среднем остается на 37% меньше данных (выводится через второй замечательный предел)

IT-Academy

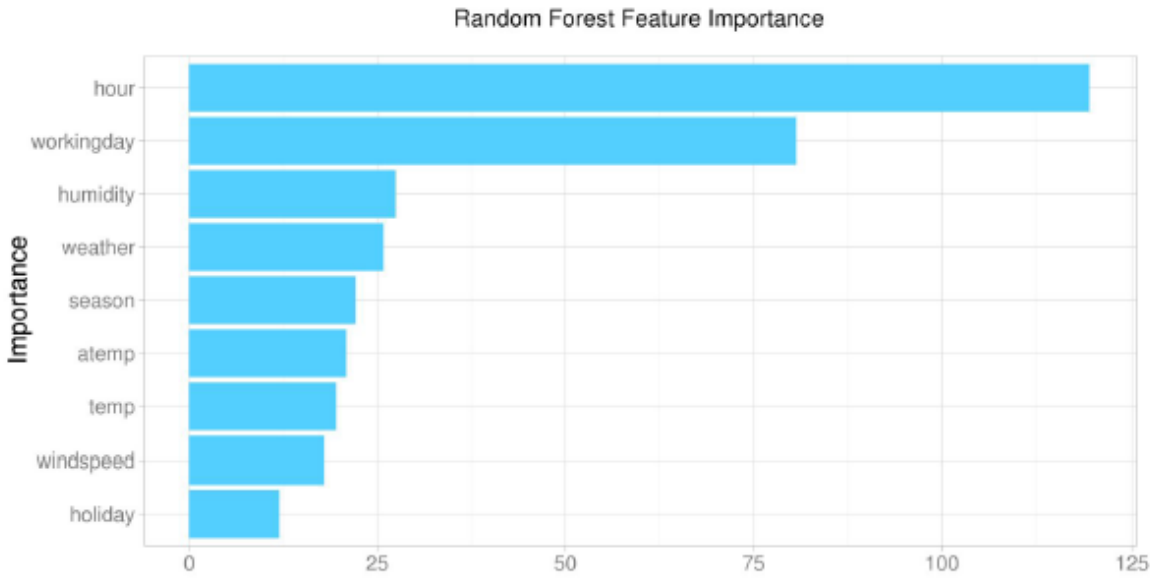
ПРАКТИКА

Bagging.ipynb

Проблем с некоторыми моделями

IT-Academy

ПРОБЛЕМА ИНТЕРПРЕТАЦИИ СЛОЖНЫХ МОДЕЛЕЙ



Случайный лес

IT-Academy

АЛГОРИТМ

Базовый алгоритм $a_i(x)$ – решающее дерево

1. Генерируем подвыборку с возвращением (bagging)
2. Строим на ней дерево $a_i(x)$, причем при каждом разбиении выбираем ***m*** случайных признаков (метод случайных подпространств)
3. Повторяем шаги 1-2 ***M*** раз
4. Усредняем (регрессия) или проводим голосование среди ответов $a_i(x)$:

$$a(x) = \frac{1}{M} \sum_{i=1}^{M} a_i(x)$$

IT-Academy

ПАРАМЕТРЫ ДЕРЕВЬЕВ

Реализация sklearn.ensemble.RandomForestClassifier/Regressor

- **criterion** – критерий построения дерева
- **max_depth** – максимальная глубина дерева
  – обычно 10-20, больше глубина –> больше риск переобучения
- **min_samples_leaf** – минимальное число объектов в листе
  – обычно 20+, больше объектов –> меньше риск переобучения

IT-Academy

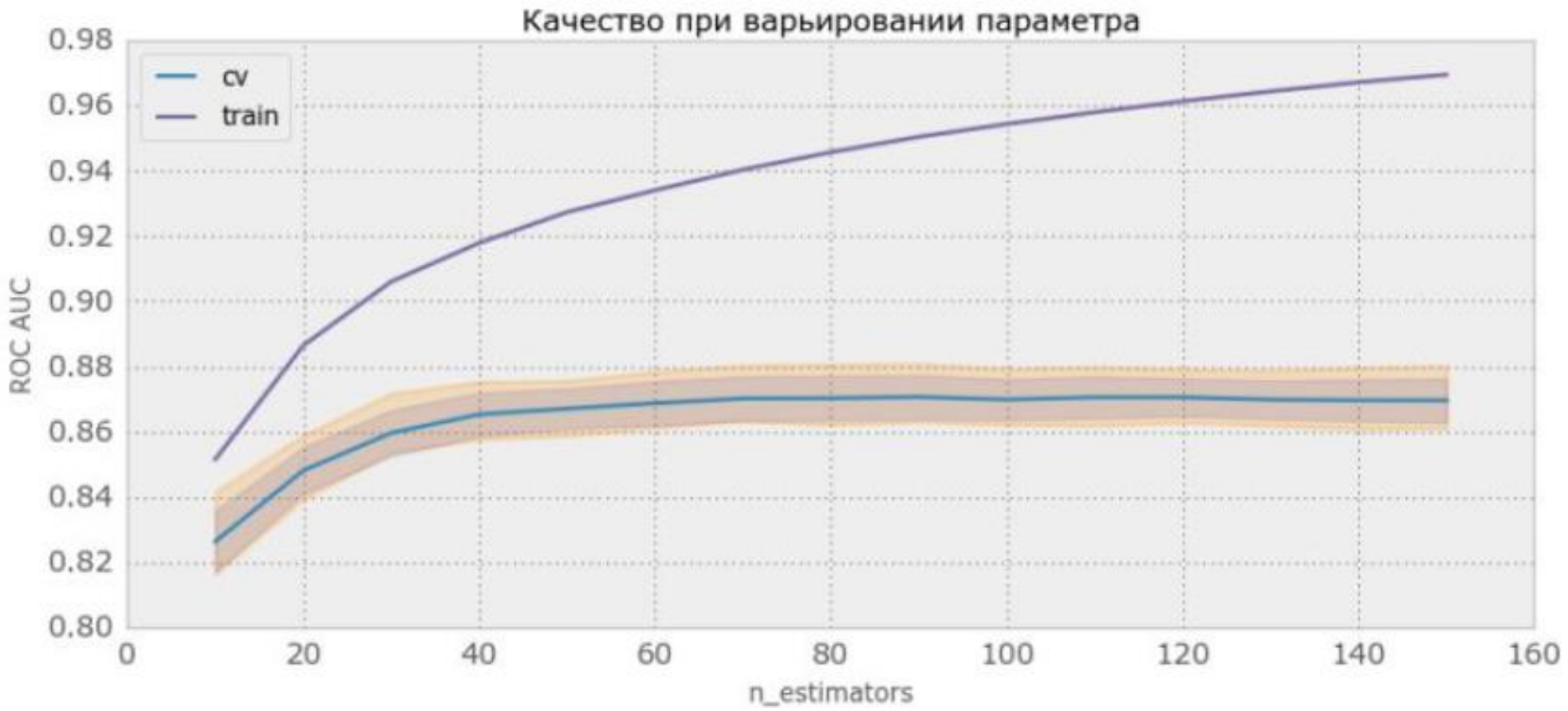
ПАРАМЕТРЫ ЛЕСА

Реализация
sklearn.ensemble.RandomForestClassifier/Regressor

- **n_estimators** - кол-во деревьев
  чем больше тем лучше
- **max_features** - число признаков случайного подпространства
- **n_jobs** - кол-во потоков для одновременного построения деревьев
  большая прибавка к скорости на многоядерных процессорах

Случайный лес

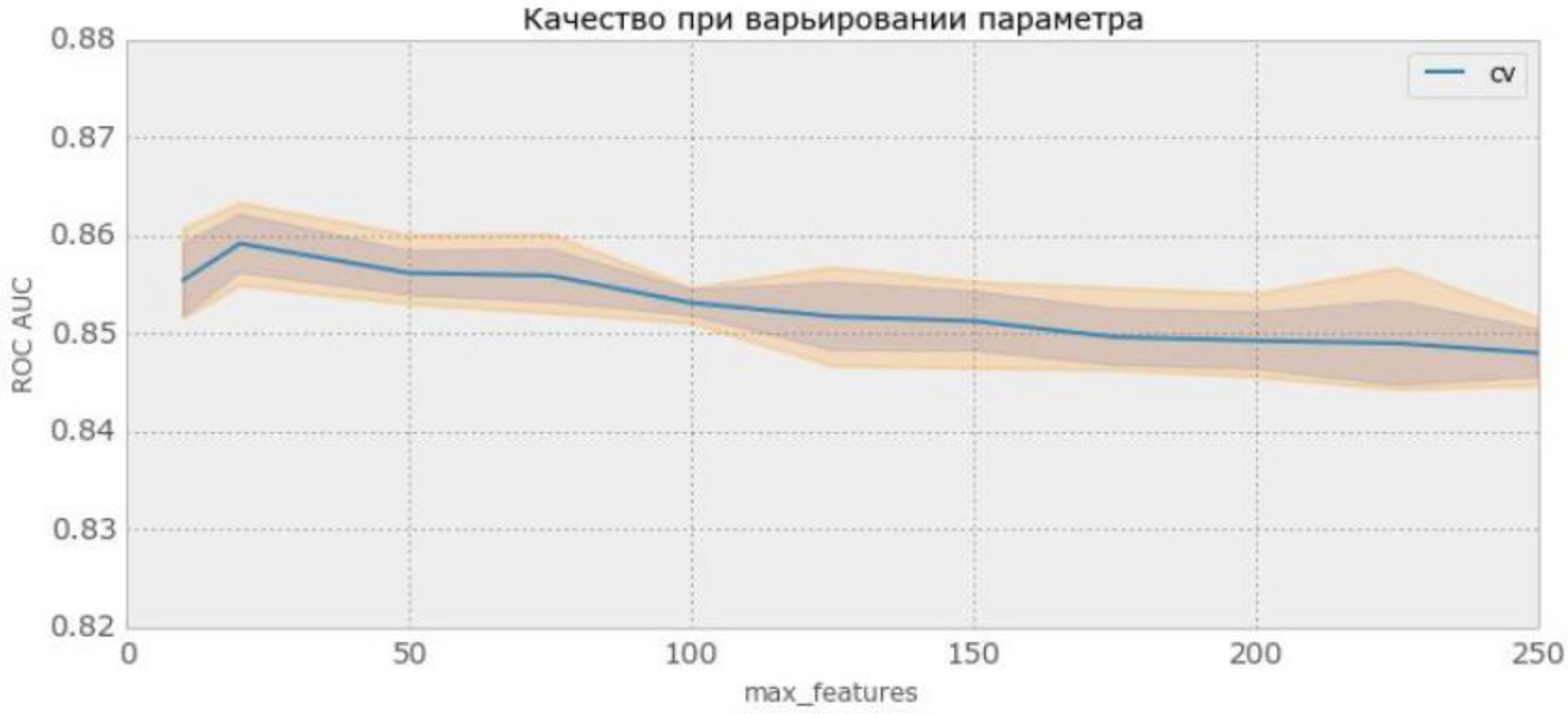
ЧИСЛО ДЕРЕВЬЕВ



Случайный лес

IT-Academy

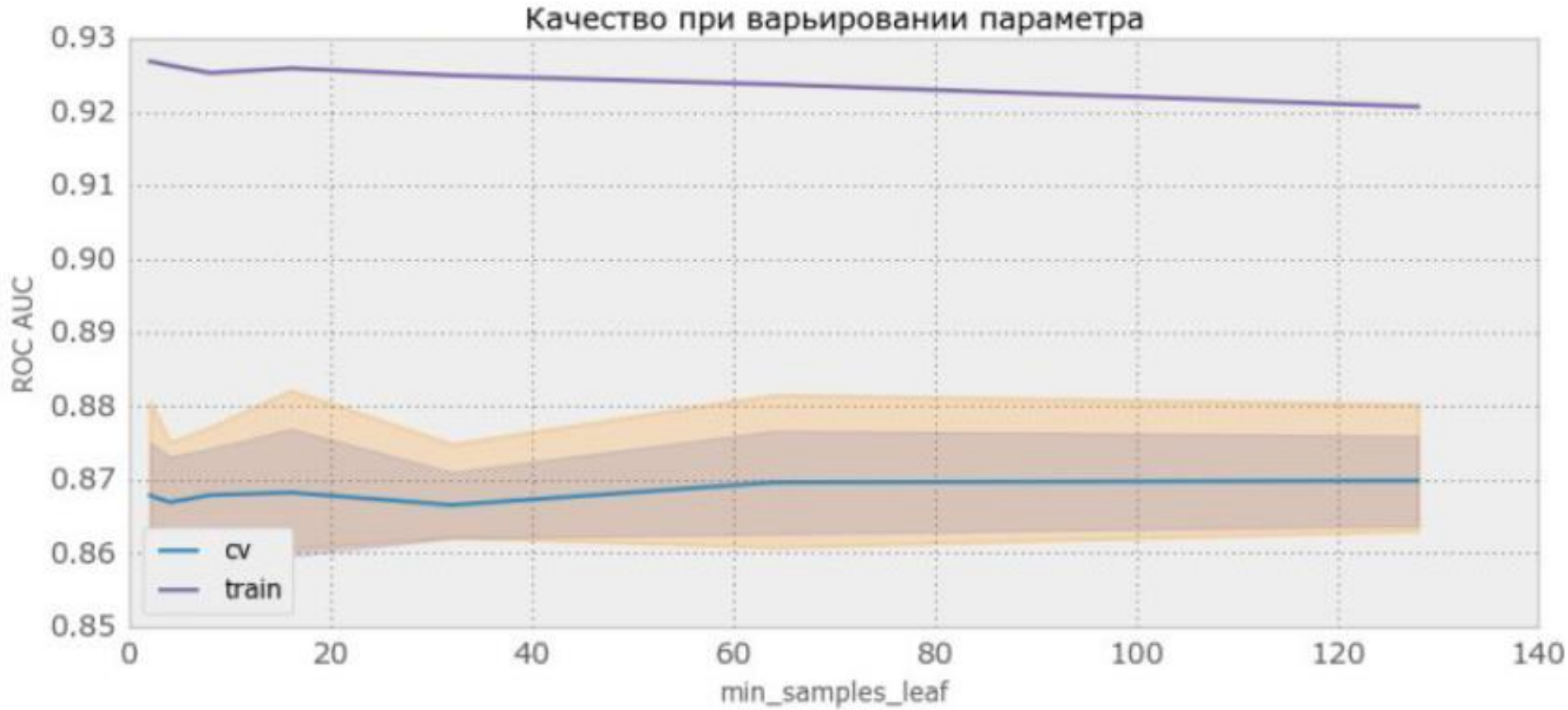
ЧИСЛО ПРИЗНАКОВ



Случайный лес

IT-Academy

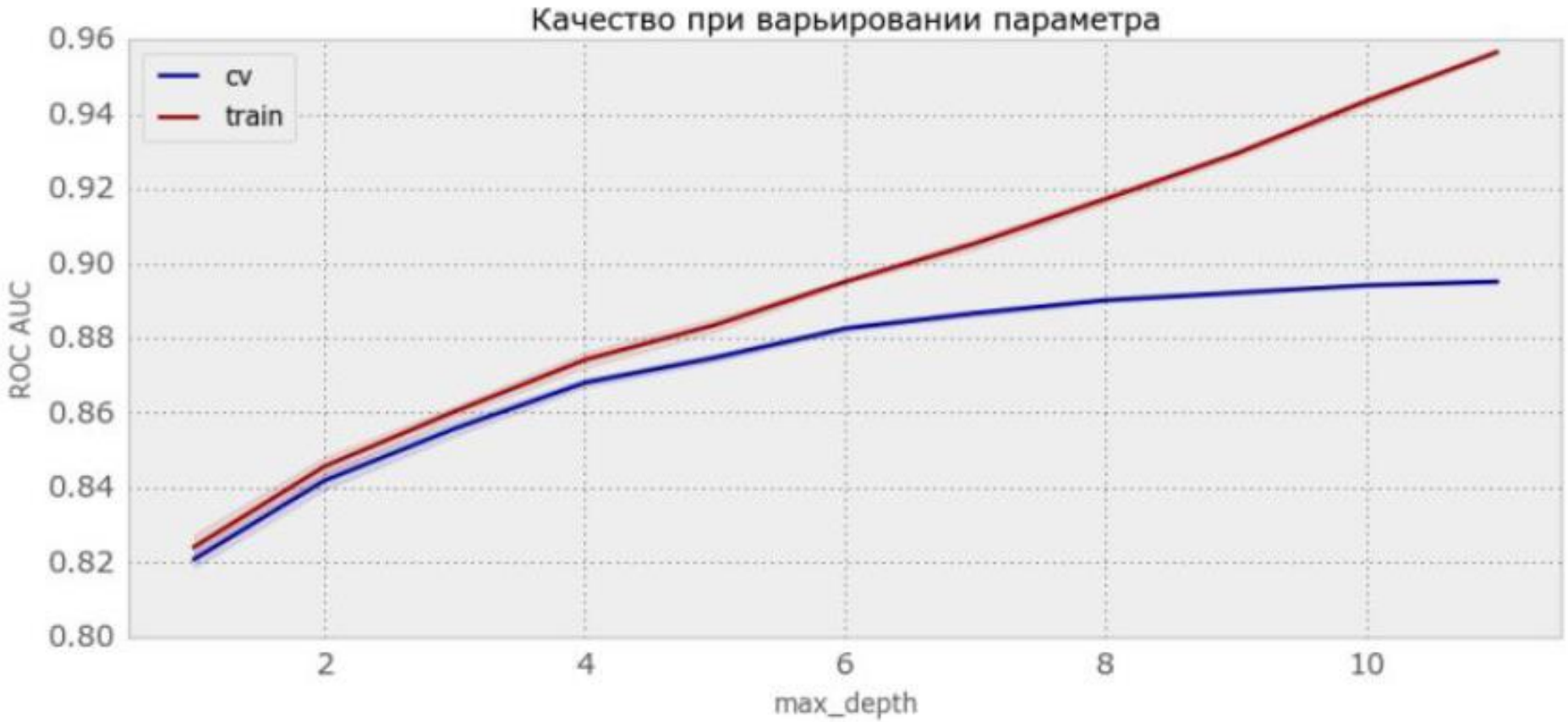
ЧИСЛО ОБЪЕКТОВ В ЛИСТЕ



Случайный лес

IT-Academy

МАКСИМАЛЬНАЯ ГЛУБИНА

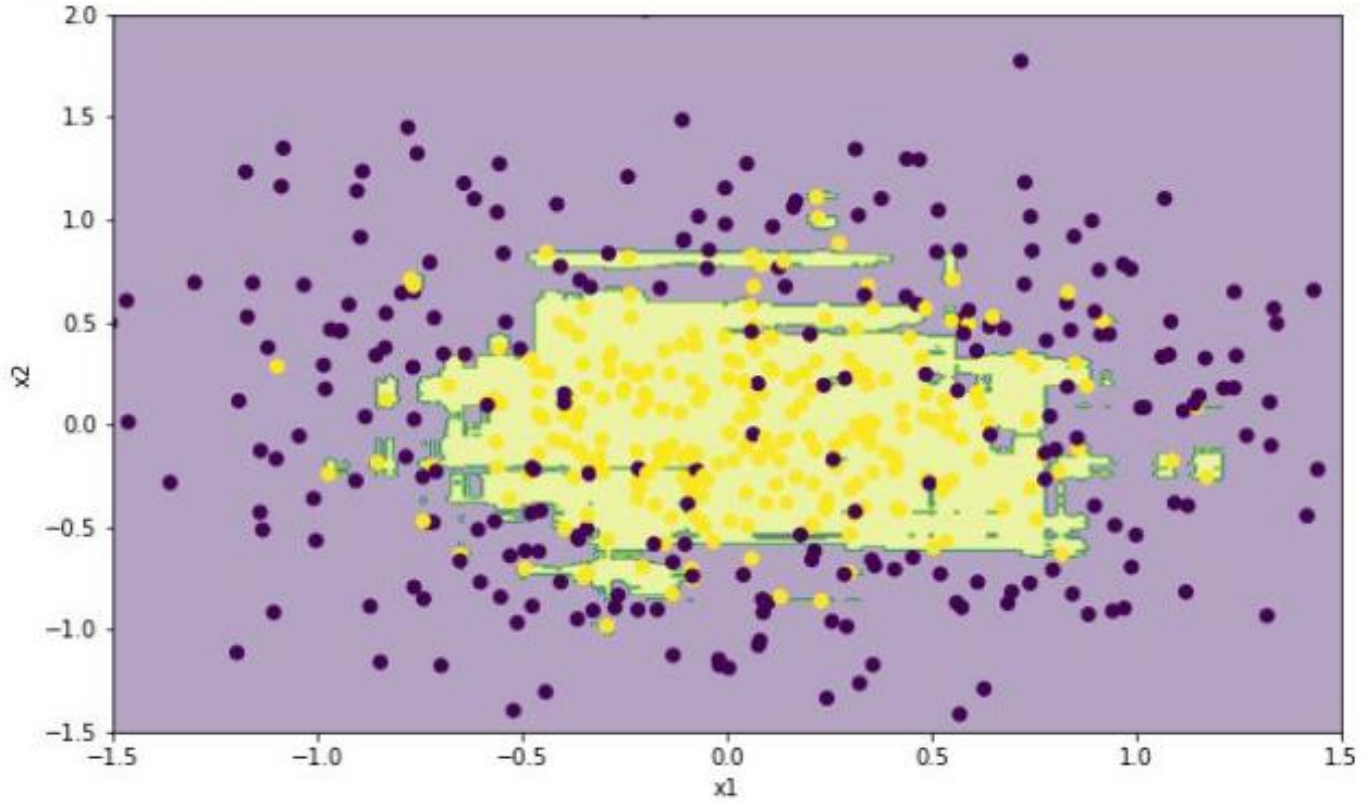


Случайный лес

IT-Academy

РАЗДЕЛЯЮЩАЯ ПОВЕРХНОСТЬ

Случайный лес



IT-Academy

ВАЖНОСТЬ ПРИЗНАКОВ

– несколько способов подсчета важности MDI, MDA, ...

– в sklearn используется MDI;

– усредненное по всем деревьям в ансамбле кол-во сплитов по признаку, взвешенное на прирост информации (Information gain) и долю объектов в вершине, в которой производится этот сплит

Случайный лес

IT-Academy

ПЛЮСЫ

- устойчив к переобучению;
- устойчивость к выбросам;
- дает хорошее качество “из коробки”;
- встроенная оценка важности признаков;
- быстрая реализация.

Случайный лес

IT-Academy

МИНУСЫ

– сложность интерпретации по сравнению с одним деревом;

– плохо справляется с очень большим числом признаков;

– работает дольше линейных моделей;

IT-Academy

ПРАКТИКА

Ensemble_methods.IPYNB

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

IT-Academy

- Как бороться с переобучением при помощи ансамблей моделей
- Как устроен Random Forest и как происходит оценка важности признаков
- Как правильно делать мета-признаки и сооружать многоуровневые модели
- Потренировались строить ансамбли разных типов на практике

IT-Academy

Ансамбли моделей. Ч1.

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ