

Машинное обучение

Лекция 8
Отбор признаков

Михаил Гуцин
mhushchyn@hse.ru

НИУ ВШЭ, 2022



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

На прошлой лекции

- ▶ Дана выборка данных X, y
- ▶ Для $k = 1 \dots K$:
 - Методом **бутстрапа** генерируем подвыборку $X^{(k)}, y^{(k)}$
 - Обучаем модель классификации или регрессии $b_k(x)$ на $X^{(k)}, y^{(k)}$
- ▶ Собираем композицию моделей:

$$\hat{y}(x) = \frac{1}{K} \sum_{k=1}^K b_k(x)$$

План

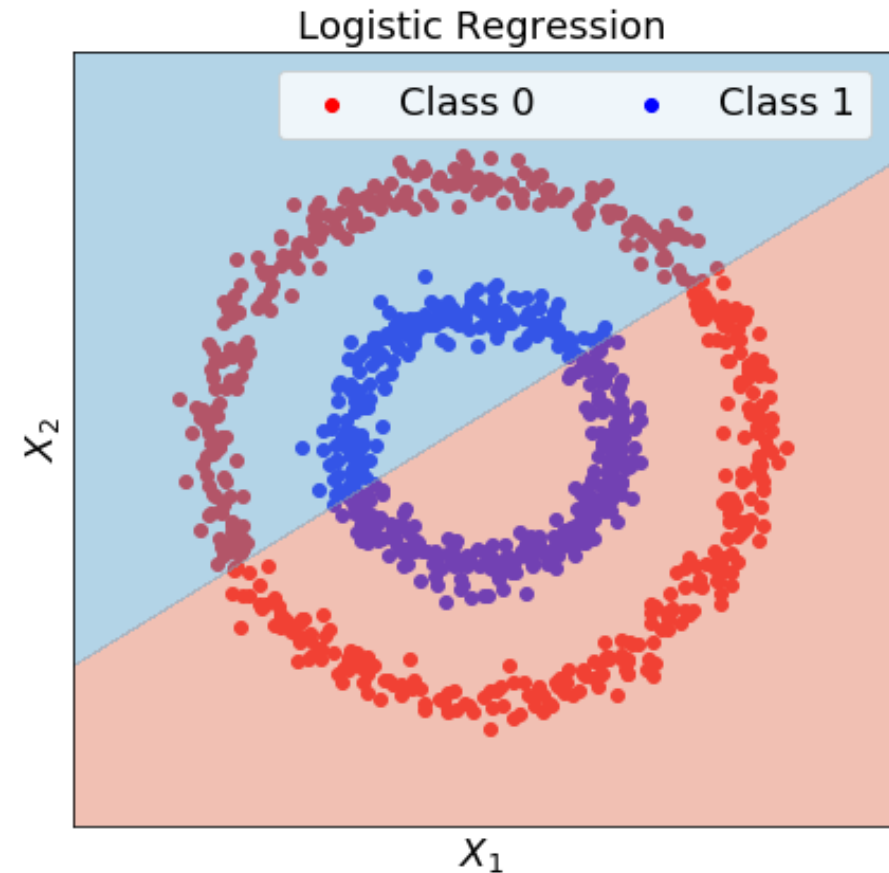
- ▶ Создание признаков
- ▶ Важность признаков
- ▶ Отбор признаков

Создание признаков (Feature engineering)



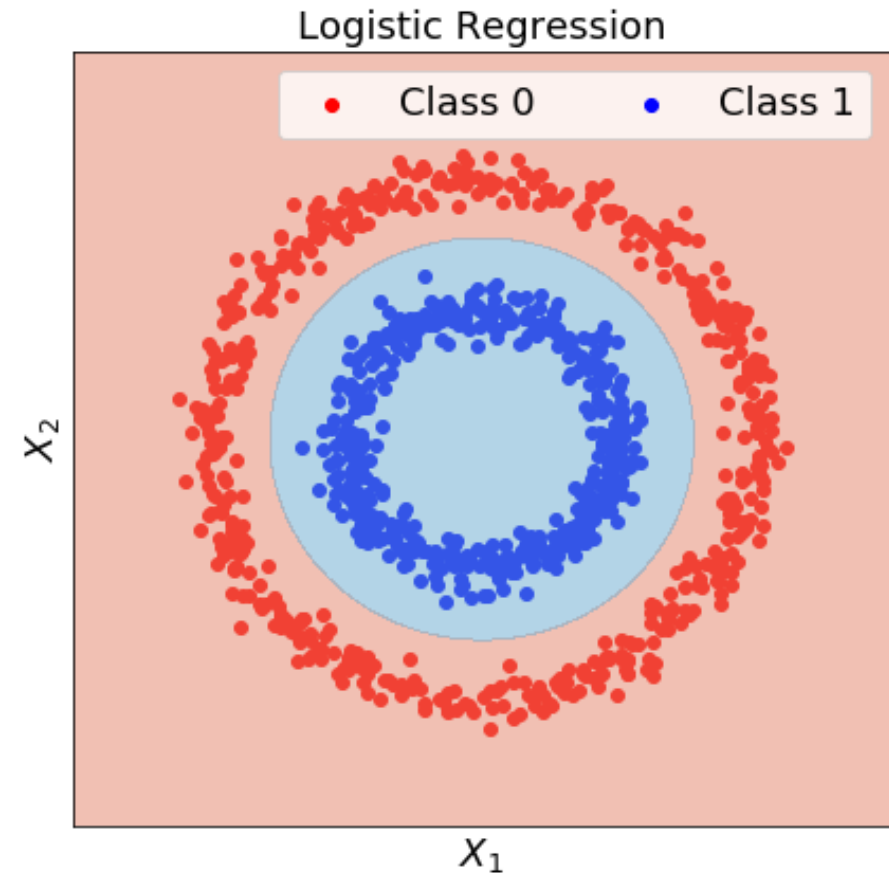
Пример 1

- ▶ Рассмотрим задачу бинарной классификации в 2D методом логистической регрессии
- ▶ Классификатор не может разделить классы используя признаки X_1 и X_2



Пример 1

- ▶ Создадим новый признак X_3 :
$$X_3 = X_1^2 + X_2^2$$
- ▶ Он позволяет разделить классы прямой линией
- ▶ Теперь логистическая регрессия решает задачу идеально используя признаки X_1 , X_2 и X_3

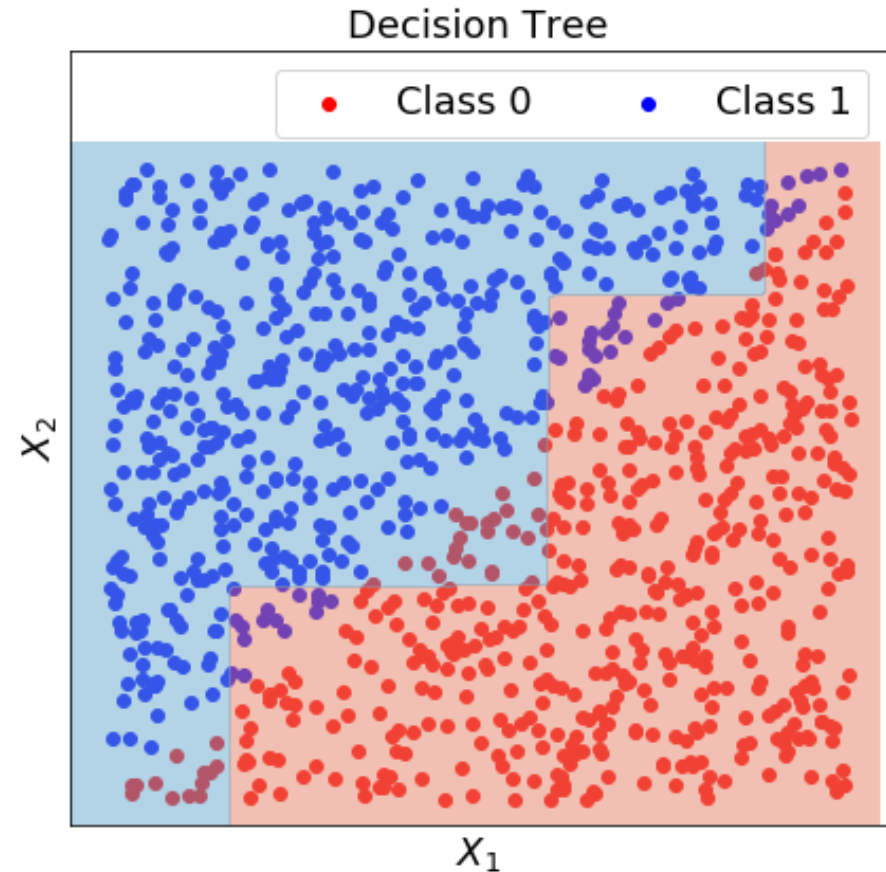


Пример 2

- ▶ Рассмотрим пример, где классы разделяются поверхностью:

$$X_2 - X_1 = 0$$

- ▶ Такая поверхность трудная для решающих деревьев
- ▶ Требуется большая глубина дерева, чтобы разделить классы

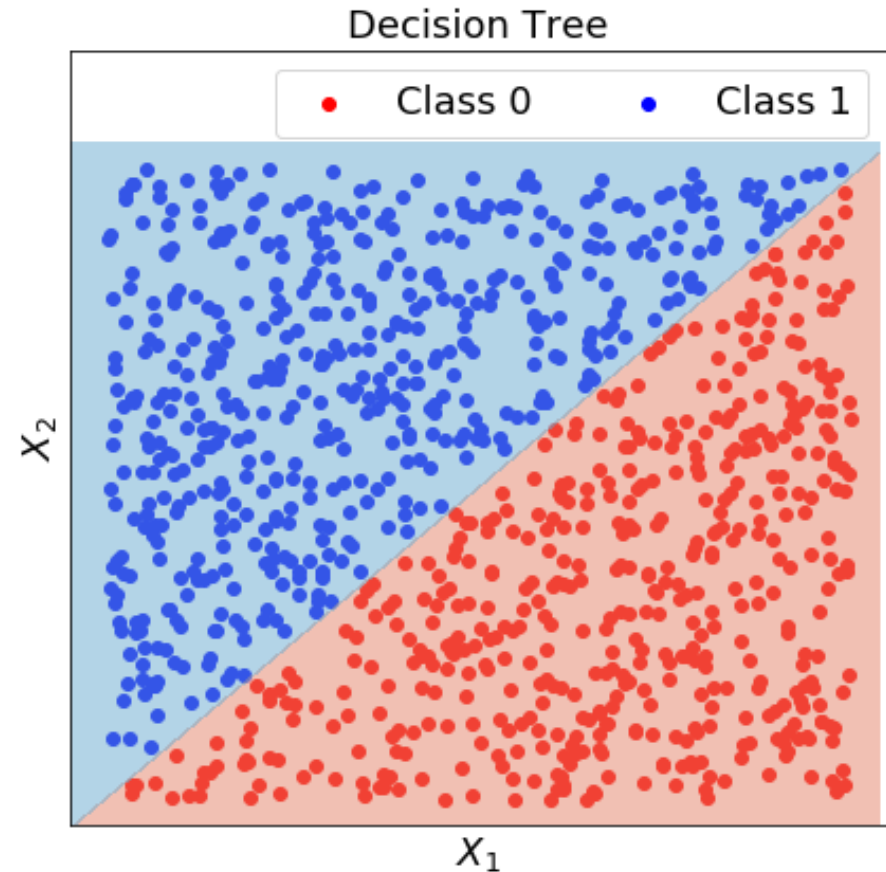


Пример 2

- ▶ Создадим новый признак X_3 , который поможет дереву:

$$X_3 = X_2 - X_1$$

- ▶ Теперь для решения задачи нужен только один предикат: $X_3 > 0$
- ▶ Требуется дерево глубины 1, чтобы решить задачу идеально 😊



Преимущества

Создание новых признаков позволяет:

- ▶ Улучшить качество моделей
- ▶ Снизить сложность моделей
- ▶ Ускорить обучение моделей
- ▶ Уменьшить размерность задачи, исключив менее информативные признаки (X_1, X_2 в примерах)

Принципы

Основные принципы создания новых признаков:

- ▶ Используйте любую информацию о задаче (классы в форме окружностей)
- ▶ Создавайте признаки со смыслом ($\sqrt{X_1^2 + X_2^2}$ - радиус)
- ▶ Компенсируйте ограничения моделей (как для дерева в примере 2)

Типичные примеры

Наиболее популярные комбинации:

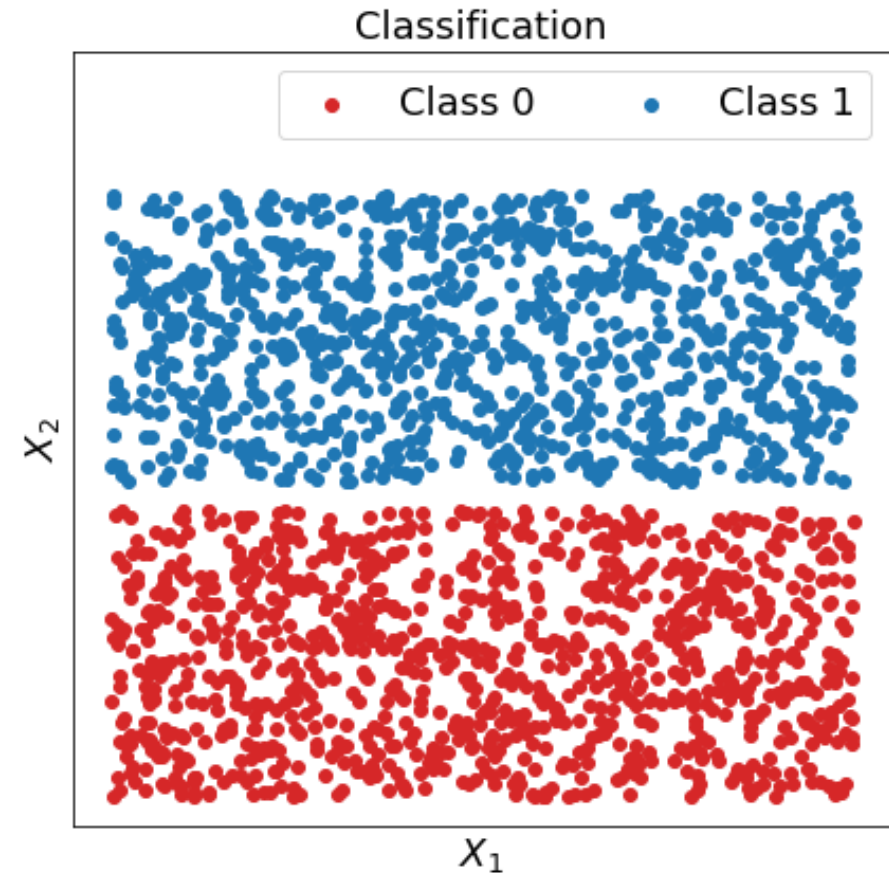
- ▶ X_i^p
- ▶ $X_1 X_2$
- ▶ $X_1^2 \pm X_2^2$
- ▶ $X_1 \pm X_2$
- ▶ $\frac{X_1 \pm X_2}{X_1 \mp X_2}$
- ▶ $\sin X_1, \cos X_1$

Важность признаков (Feature importance)



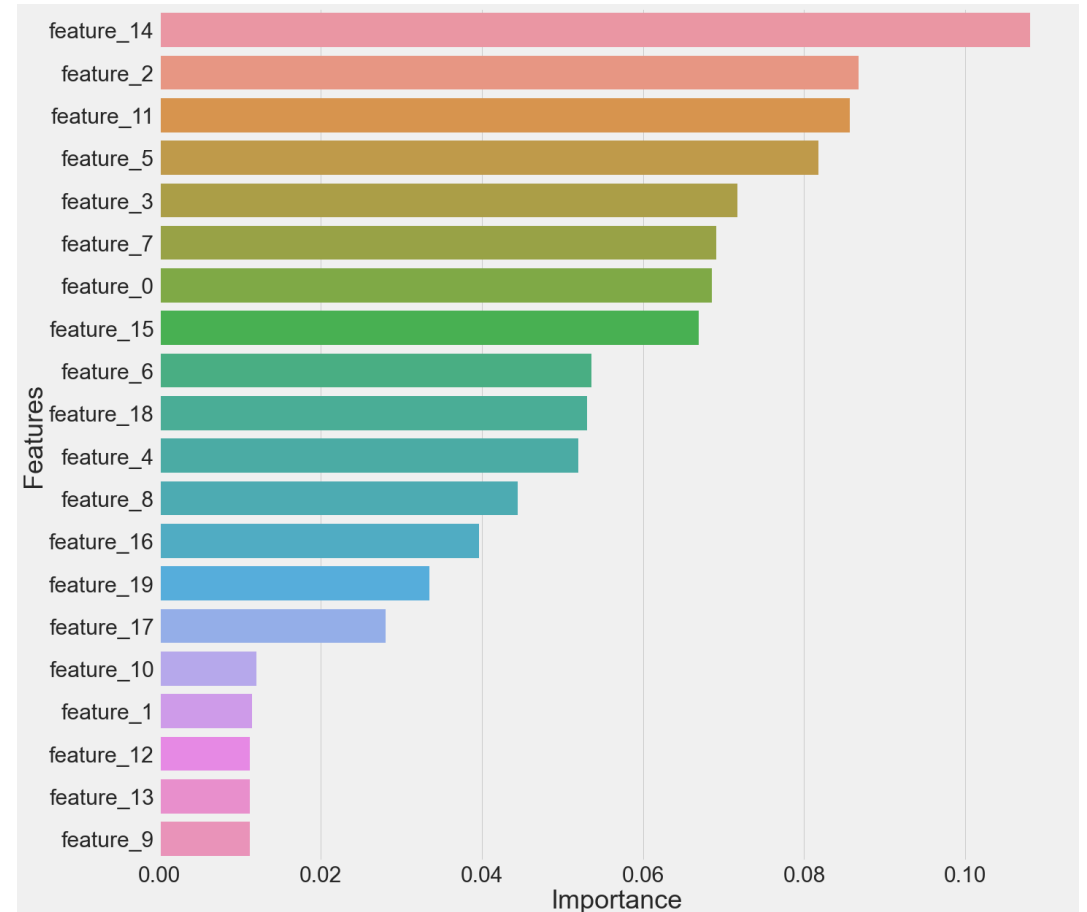
Интуиция

- ▶ Не все признаки одинаково полезны для решения задачи
- ▶ Некоторые из них более информативны, чем другие
- ▶ Например, X_1 неинформативна для классификации
- ▶ **Цель** – определить **важность каждого признака**



Методы решения

- ▶ Подсчет корреляций
- ▶ Вероятностные критерии
- ▶ Решающие деревья
- ▶ Линейные модели
- ▶ Общий метод



Корреляции

Для признака f вычислим его корреляцию с целевой переменной y :

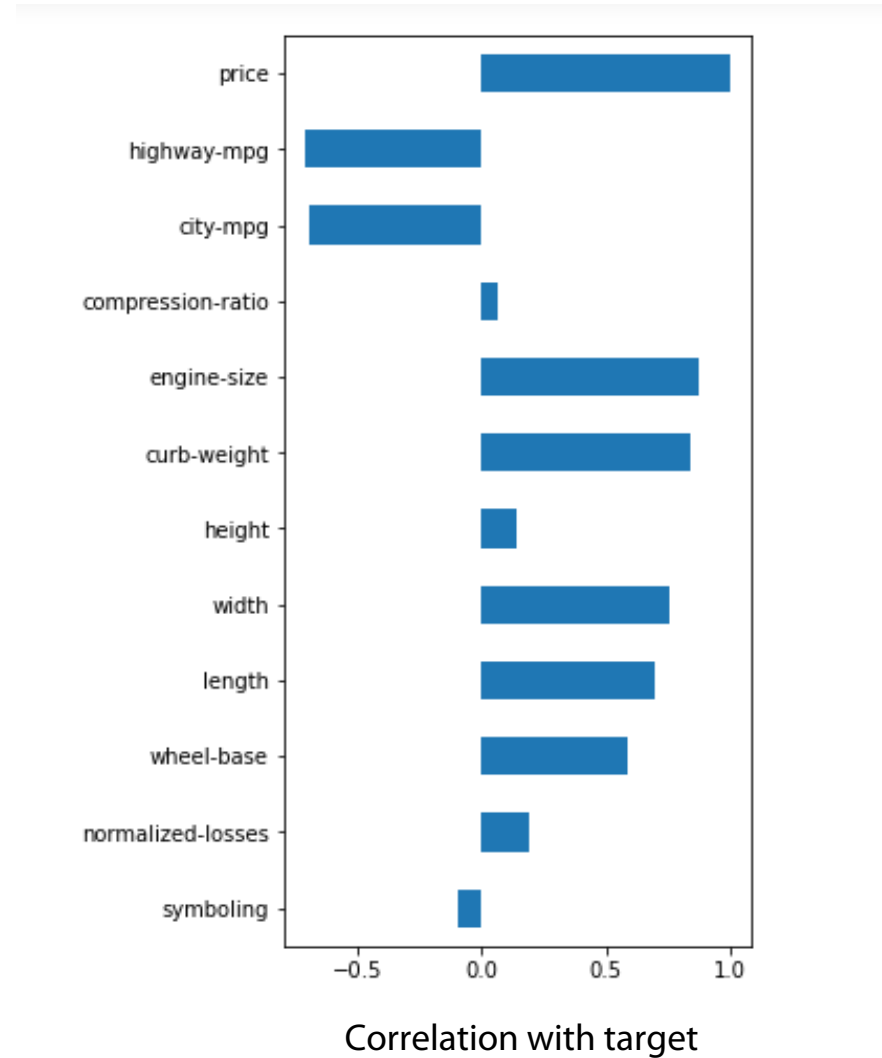
$$\rho(f, y) = \frac{\sum_i (f_i - \bar{f})(y_i - \bar{y})}{\sqrt{\sum_i (f_i - \bar{f})^2 \sum_i (y_i - \bar{y})^2}}$$

y_i - метки в классификации или целевая переменная в регрессии для i -го объекта

f_i - значение признака для i -го объекта

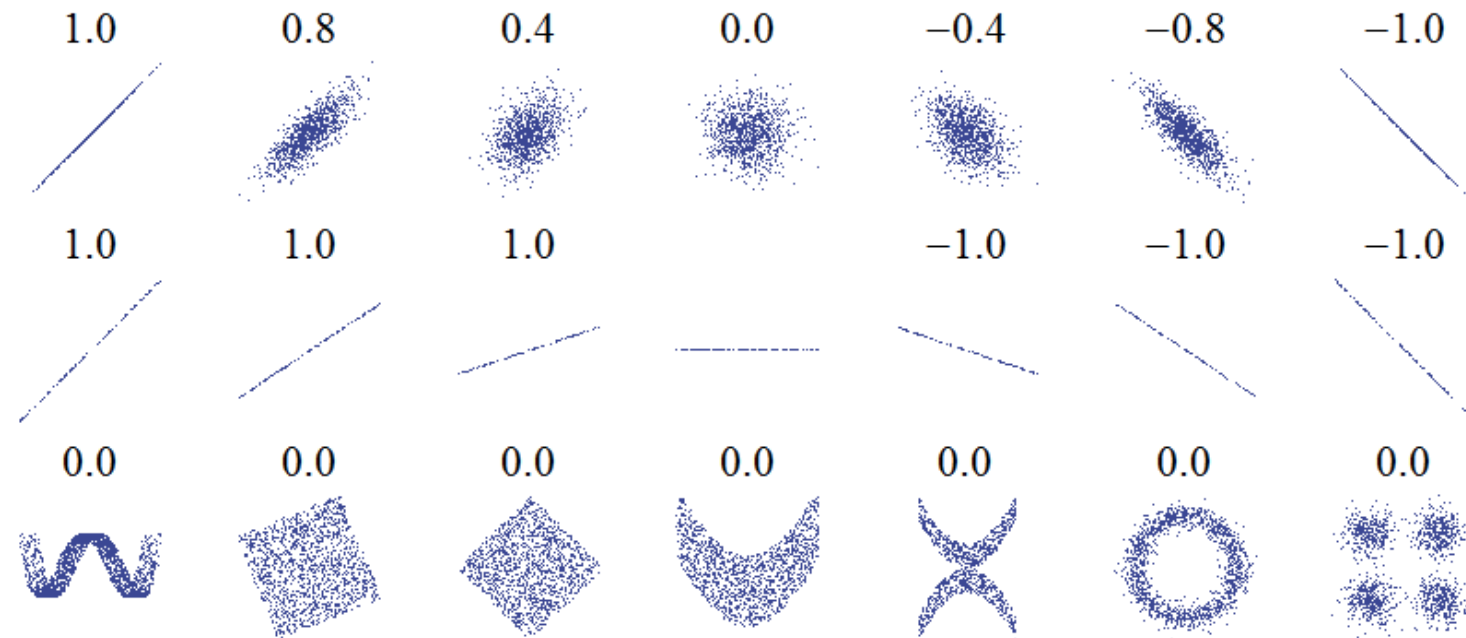
$\rho(f, y)$ - важность признака $Imp(f)$

Пример

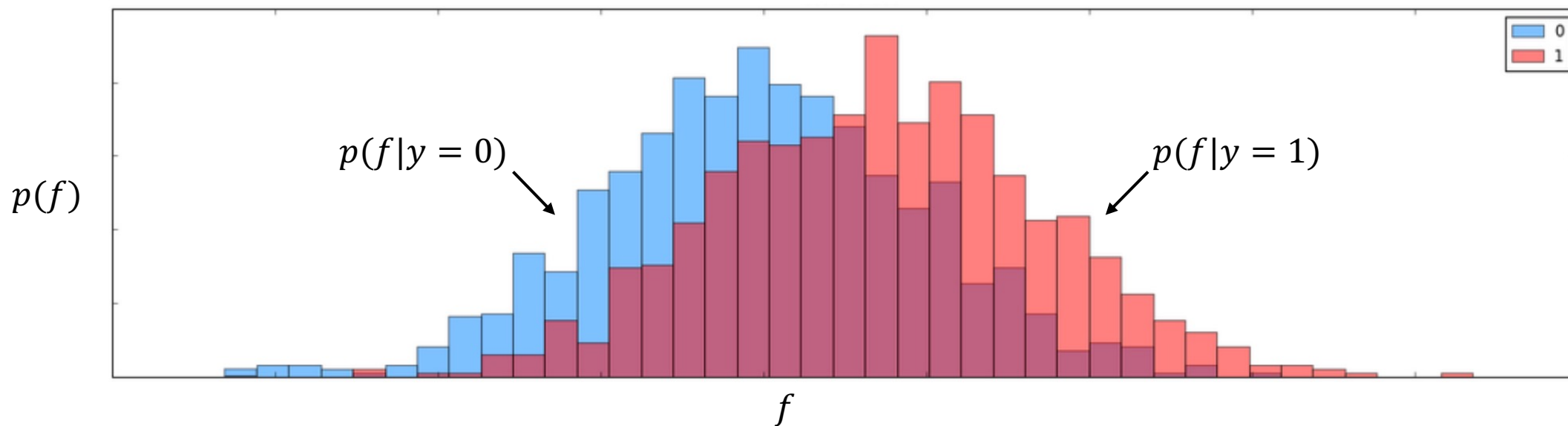


Корреляция

- ▶ Легко посчитать
- ▶ Но отражает только линейные зависимости



Вероятностные критерии (расстояния)

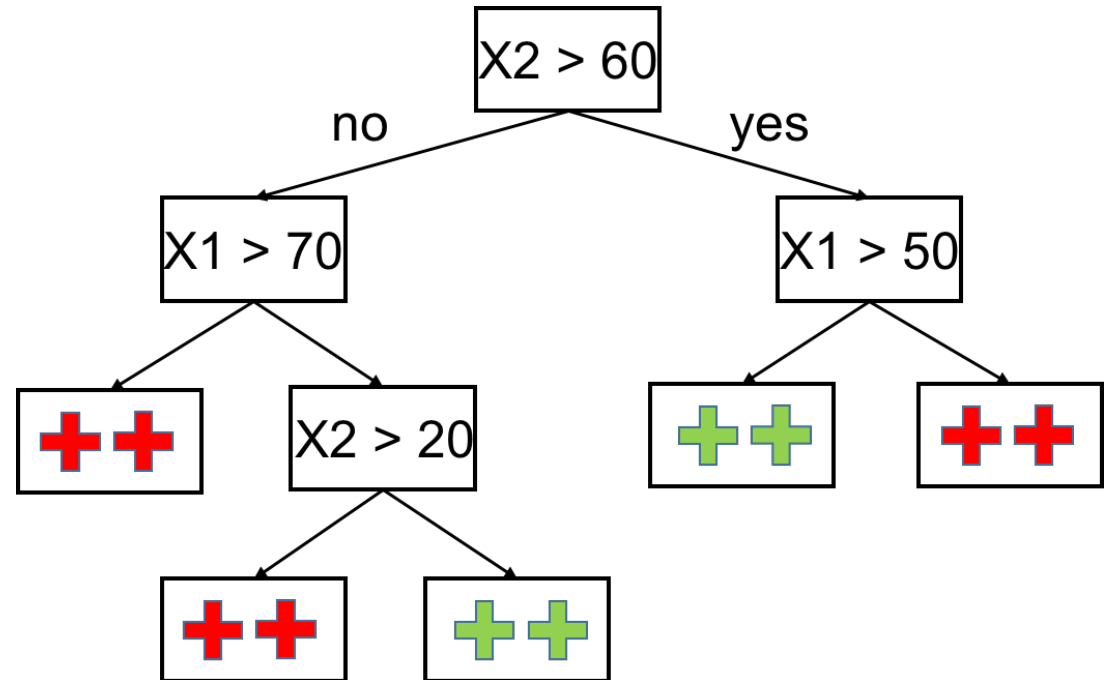


Важность признака $Imp(f)$ как расстояние полной вариации между двумя распределениями:

$$Imp(f) = \int |p(f|y=1) - p(f|y=0)|df$$

Решающие деревья

- ▶ Каждая вершина t имеет двух потомков
- ▶ n_t - число объектов в вершине
- ▶ $I(t)$ – значения критерия информативности (gini, cross-entropy, MSE) для вершины



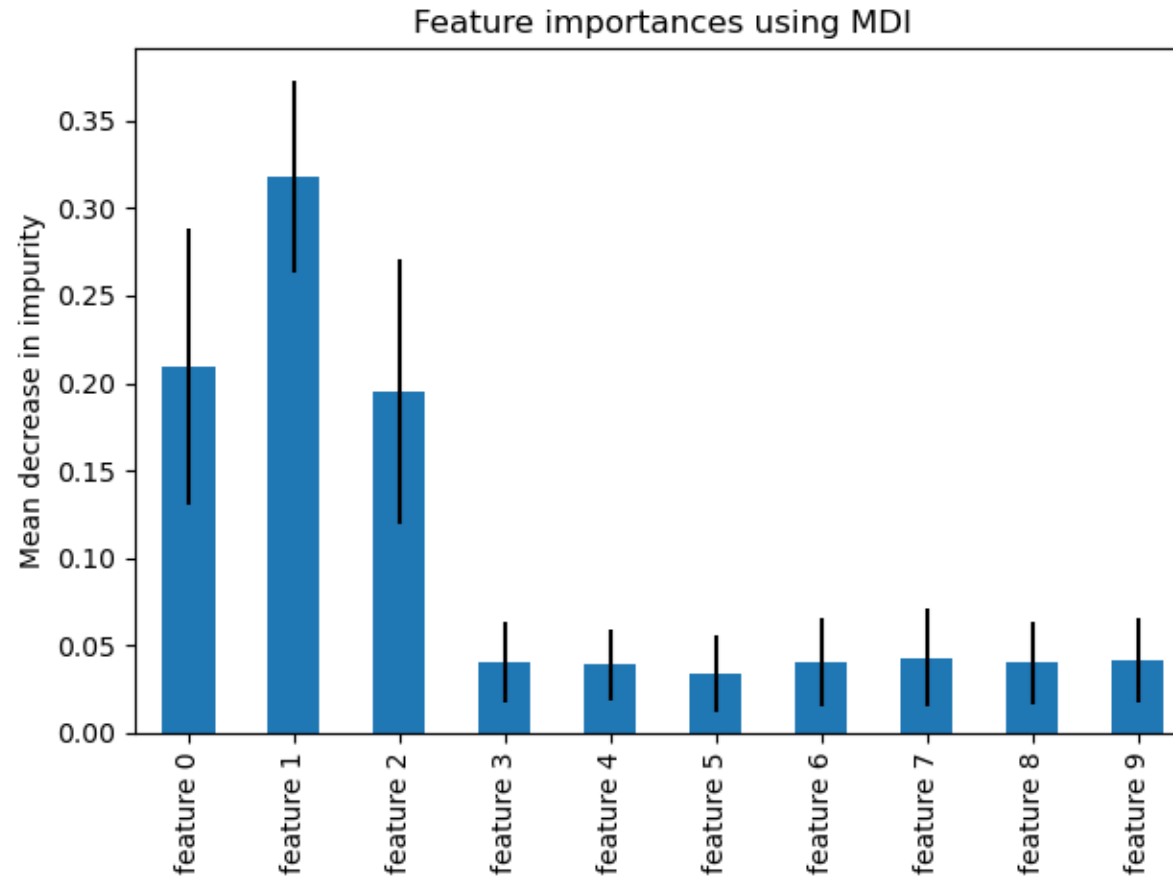
Решающие деревья

Пусть $T(f)$ – набор всех вершин дерева, где использовался признак f в предикате. Тогда важность признака f :

$$Imp(f) = \sum_{t \in T(f)} n_t \Delta I(t)$$

$$\Delta I(t) = I(t) - \sum_{c \in children(t)} \frac{n_c}{n_t} I(c)$$

Пример



https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Линейные модели

Рассмотрим линейную модель с регуляризацией (L_1 или L_2):

$$\hat{y} = w_0 + w_1 f_1 + w_2 f_2 + \dots + w_k f_k$$

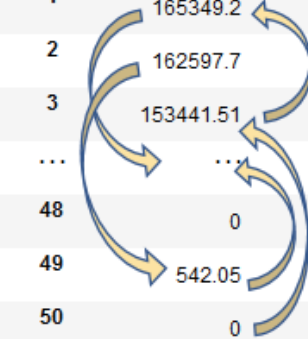
Если признаки нормированы (значения одного масштаба), то важность признака f_i равна:

$$Imp(f_i) = |w_i|$$

Общий метод

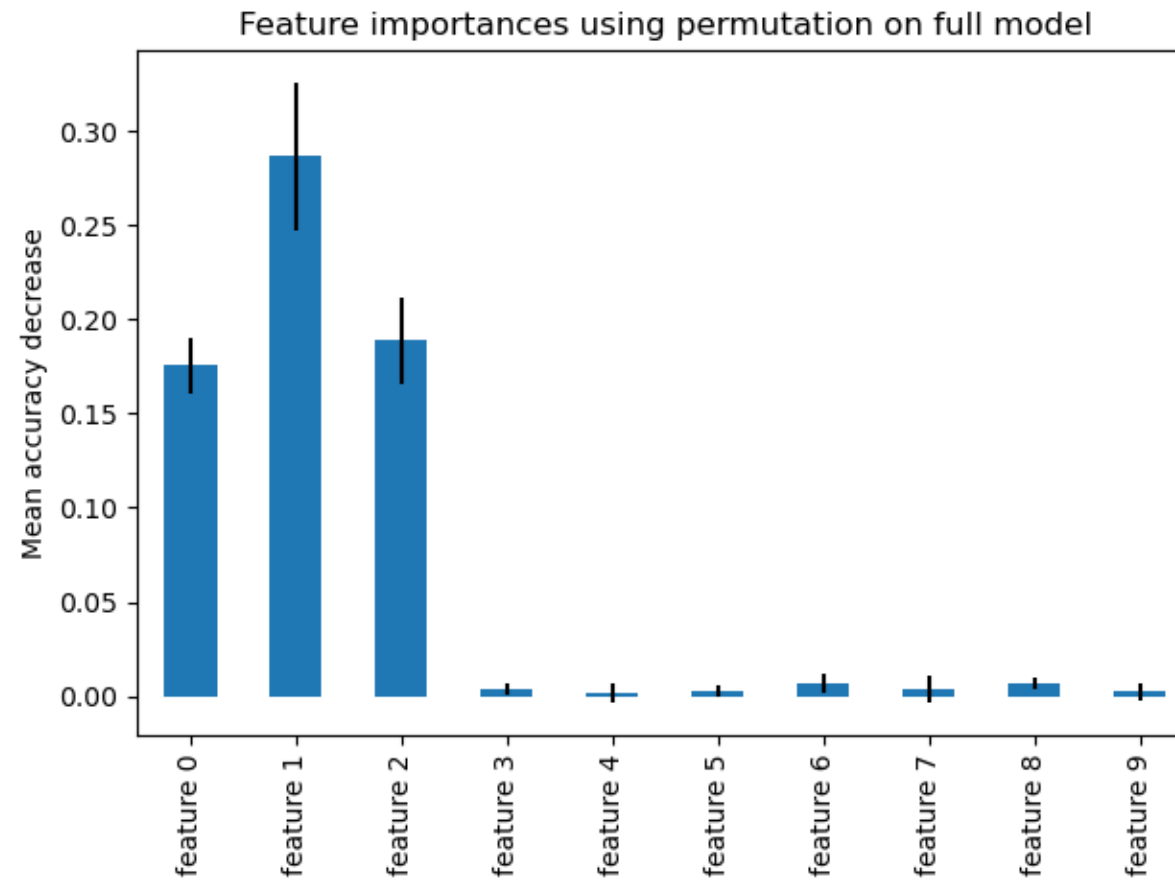
- ▶ Обучаем модель
- ▶ Считаем метрику качества Q_0 на тестовой выборке
- ▶ Для каждого признака f :
 - Случайно перемешиваем значения признака
 - Считаем новое значение метрики Q_f на тестовой выборке
 - Определяем важность признака:

$$Imp(f) = Q_0 - Q_f$$



	RD Spend	Administration	Marketing Spend	Profit	state_California
1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

Пример

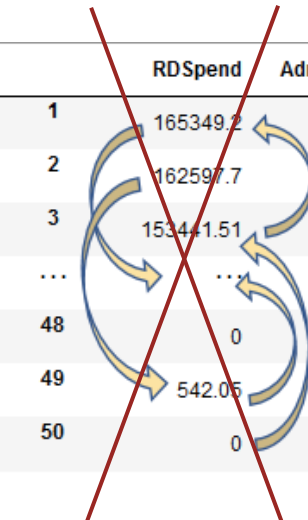


https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Общий метод (модификация)

- ▶ Обучаем модель на всем наборе признаков
- ▶ Считаем метрику качества Q_0 на тестовой выборке
- ▶ Для каждого признака f :
 - Обучаем модель без этого признака
 - Считаем новое значение метрики Q_f на тестовой выборке
 - Определяем важность признака:

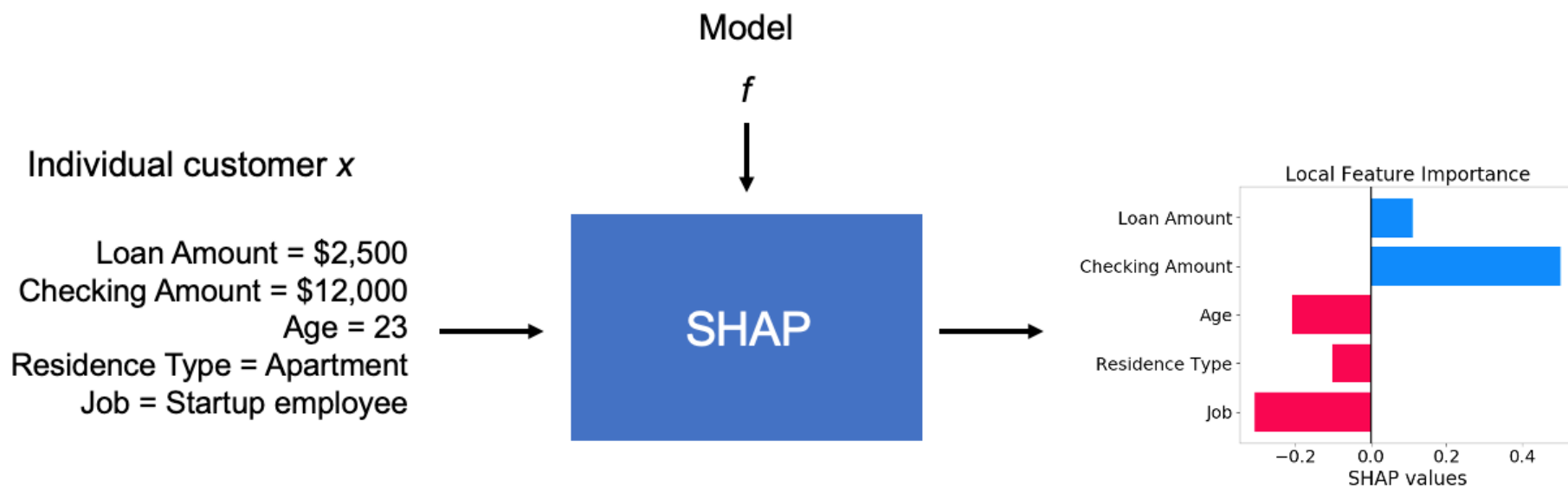
$$Imp(f) = Q_0 - Q_f$$



	RDSpend	Administration	MarketingSpend	Profit	state_California
1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

SHAP

- ▶ Существует множество других методов
- ▶ Один из наиболее интересный методов основан на векторах Шепли (Shapley values): <https://github.com/slundberg/shap>



Отбор признаков (Feature Selection)



Отбор признаков

Цель – уменьшить число признаков с минимальными потерями качества модели.

Примеры:

- ▶ Оставить K из D наилучших признаков
- ▶ Удалить как можно больше признаков так, чтобы качество модели $Q \geq Q_{min}$

All Features



Feature Selection



Final Features



Методы

- ▶ Метод фильтрации
- ▶ Встроенные методы
- ▶ Рекурсивное удаление признаков (recursive feature elimination)

Метод фильтрации

- ▶ Определяем важность отдельных признаков:

$$Imp(f_1), Imp(f_2), \dots, Imp(f_D)$$

- ▶ Выбираем нужное число признаков с наибольшими значениями важности

- ▶ Простой и быстрый метод
- ▶ Плохо работает для скоррелированных признаков

Встроенные методы

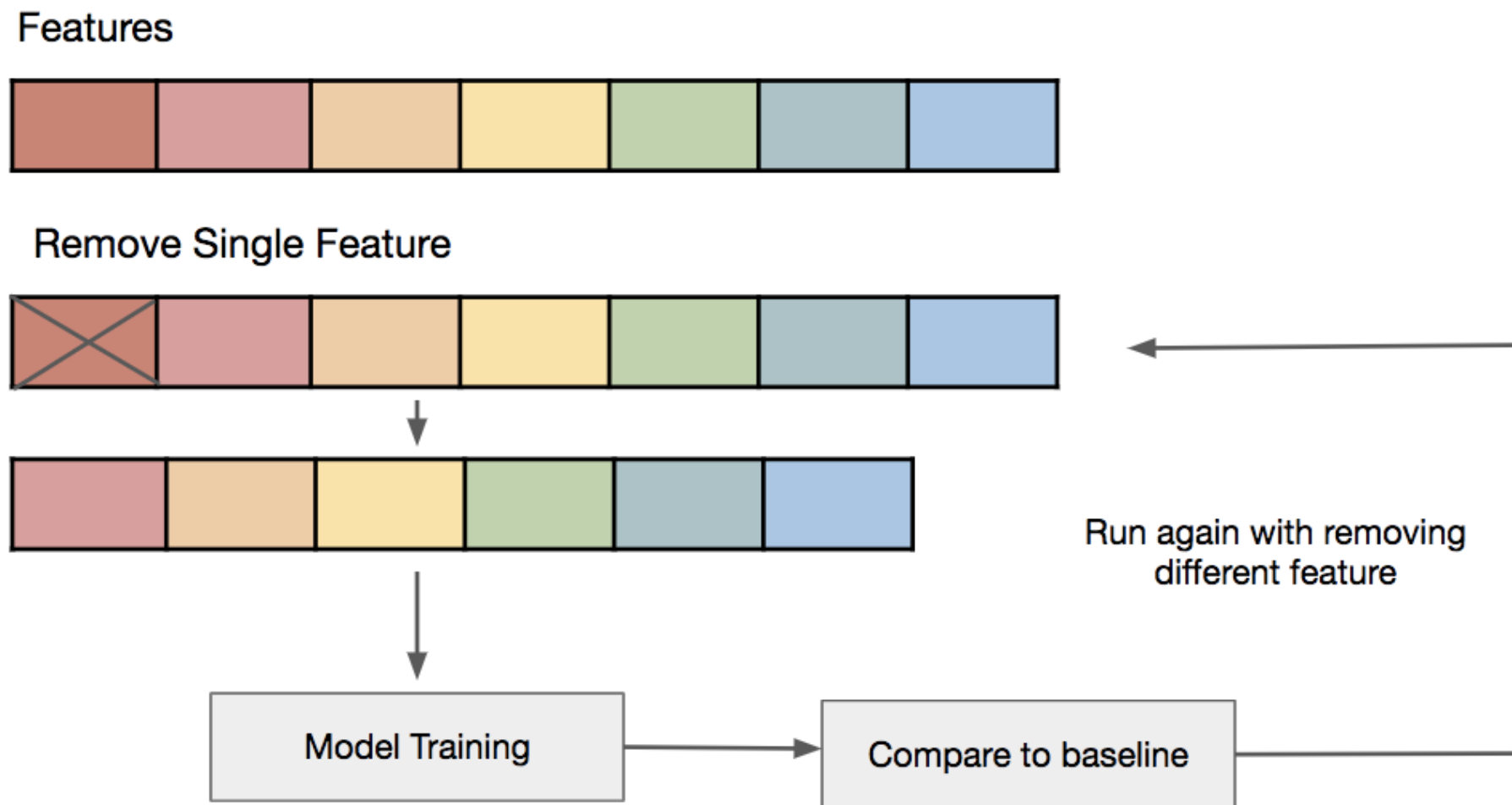
- ▶ Определяем важность отдельных признаков с помощью решающих деревьев или линейных моделей:

$$Imp(f_1), Imp(f_2), \dots, Imp(f_D)$$

- ▶ Выбираем нужное число признаков с наибольшими значениями важности

- ▶ Популярный метод
- ▶ Учитывает корреляции между признаками

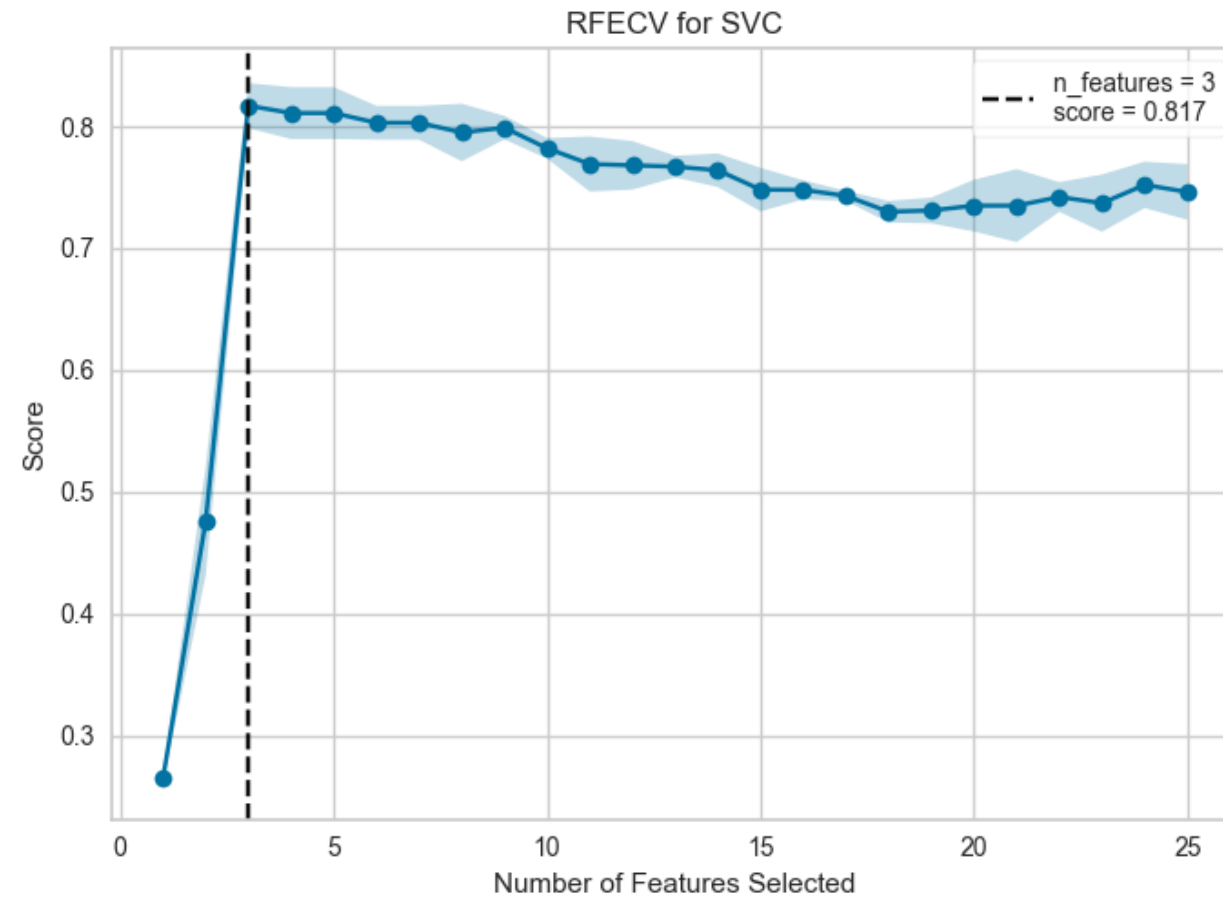
Рекурсивное удаление признаков



Рекурсивное удаление признаков

- ▶ Обучаем модель на всем наборе признаков
- ▶ Определите важность признаков используя эту модель
- ▶ Удаляем наименее важный признак
- ▶ Повторяем процедуру

Пример



Заключение

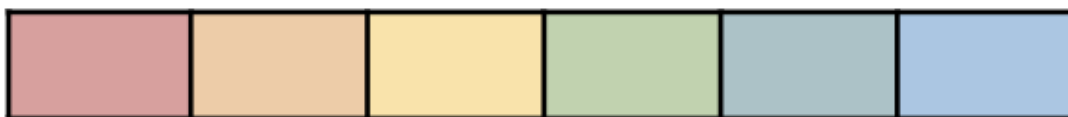


Резюме

Features



Remove Single Feature



Model Training

Compare to baseline

Run again with removing
different feature

