

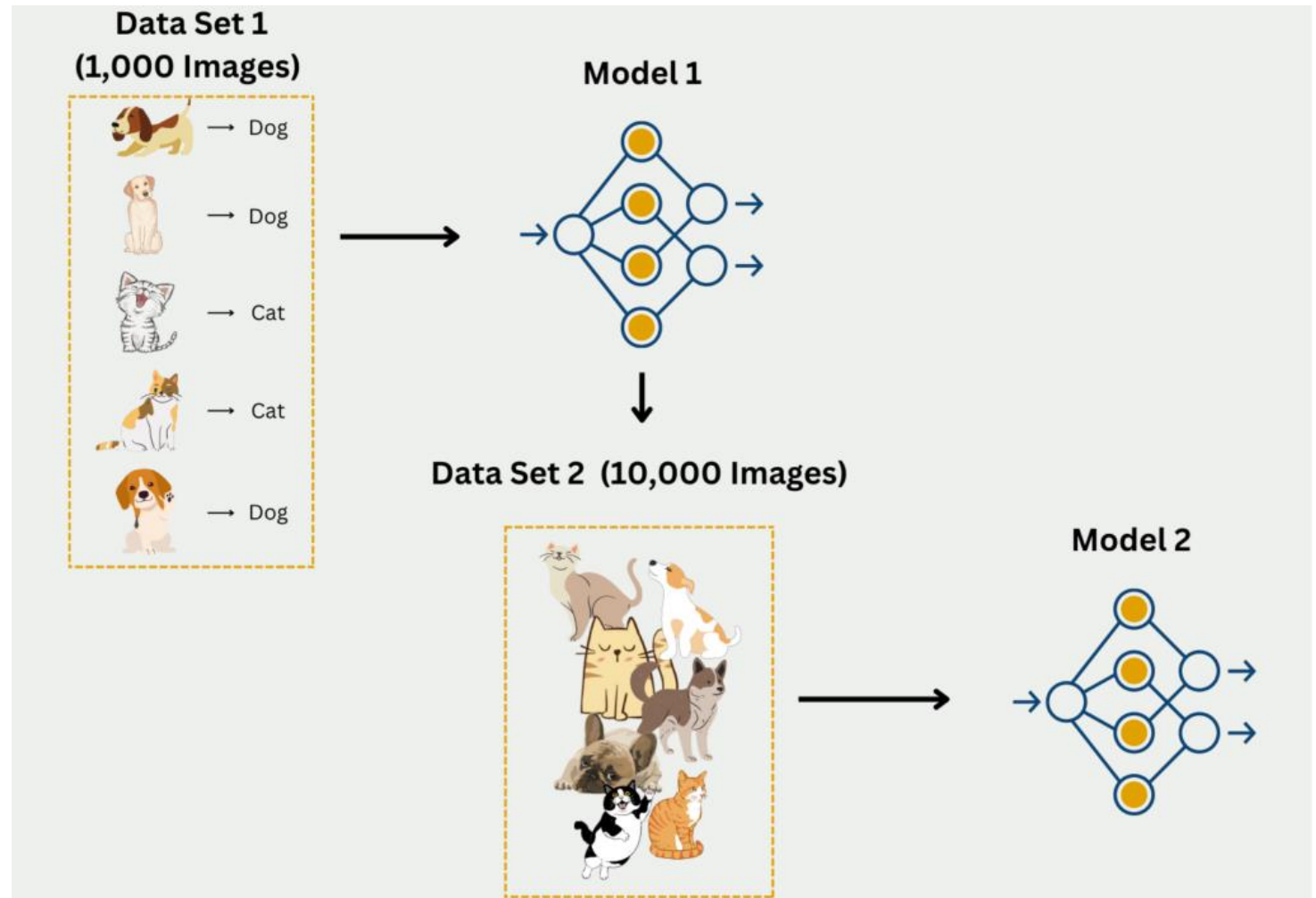
Дополнительная лекция

Лекция 10

Semi-supervised learning

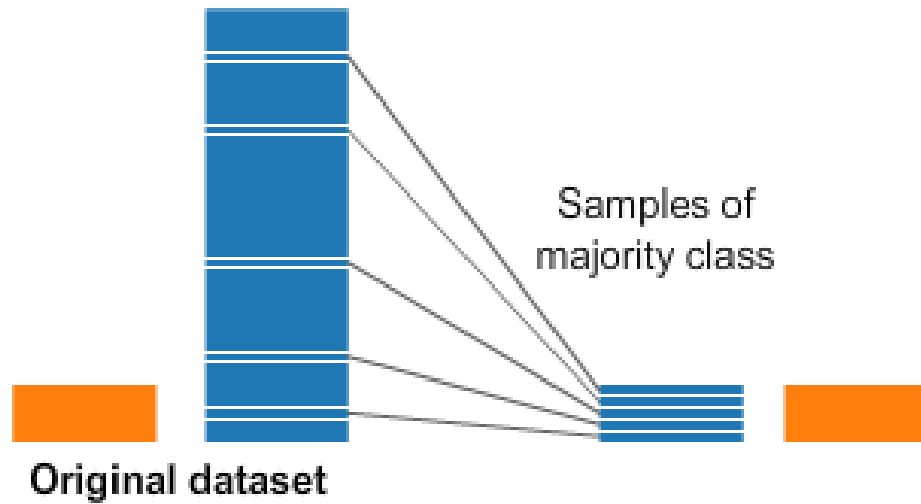
Не всегда есть
разметка на всем
множестве
наблюдений

Можно обучить
**промежуточную
модель**, чтобы
доразметить
датасет



Борьба с дисбалансом

Undersampling



Oversampling

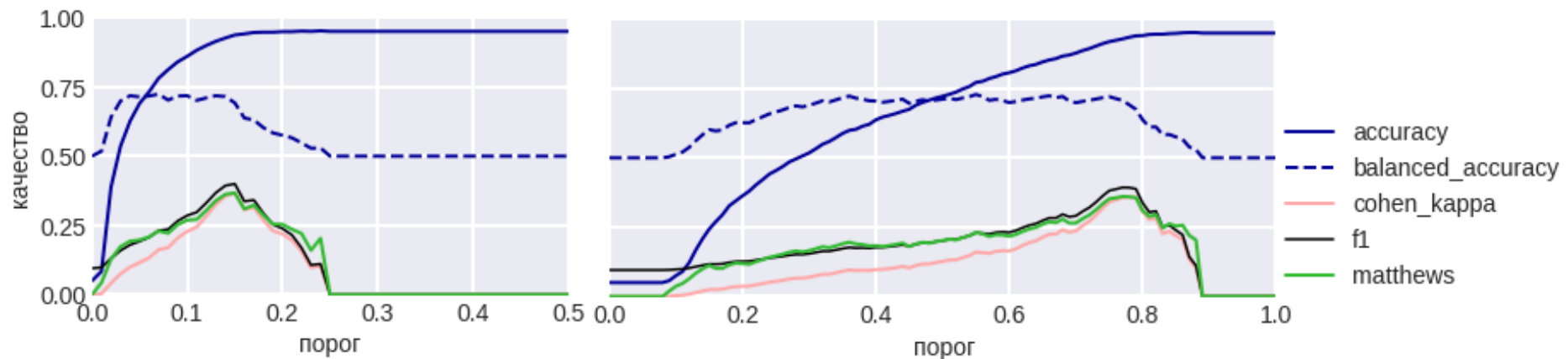


Борьба с дисбалансом

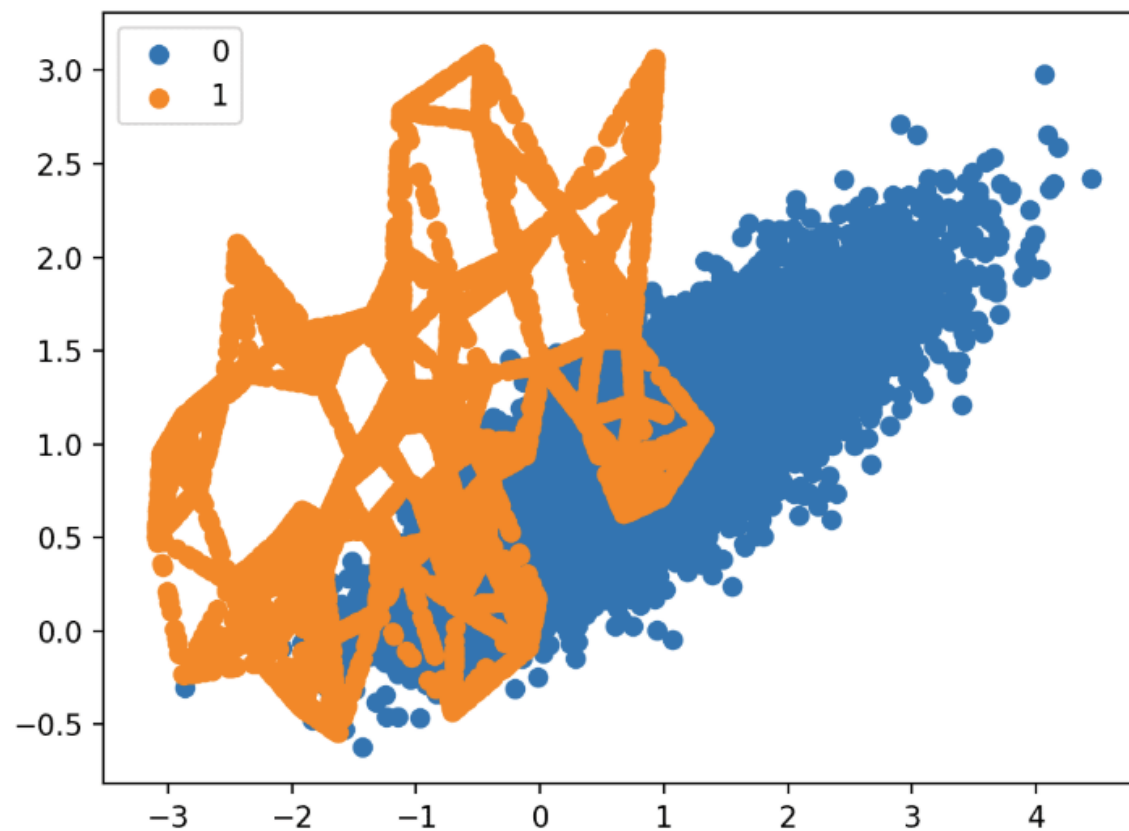
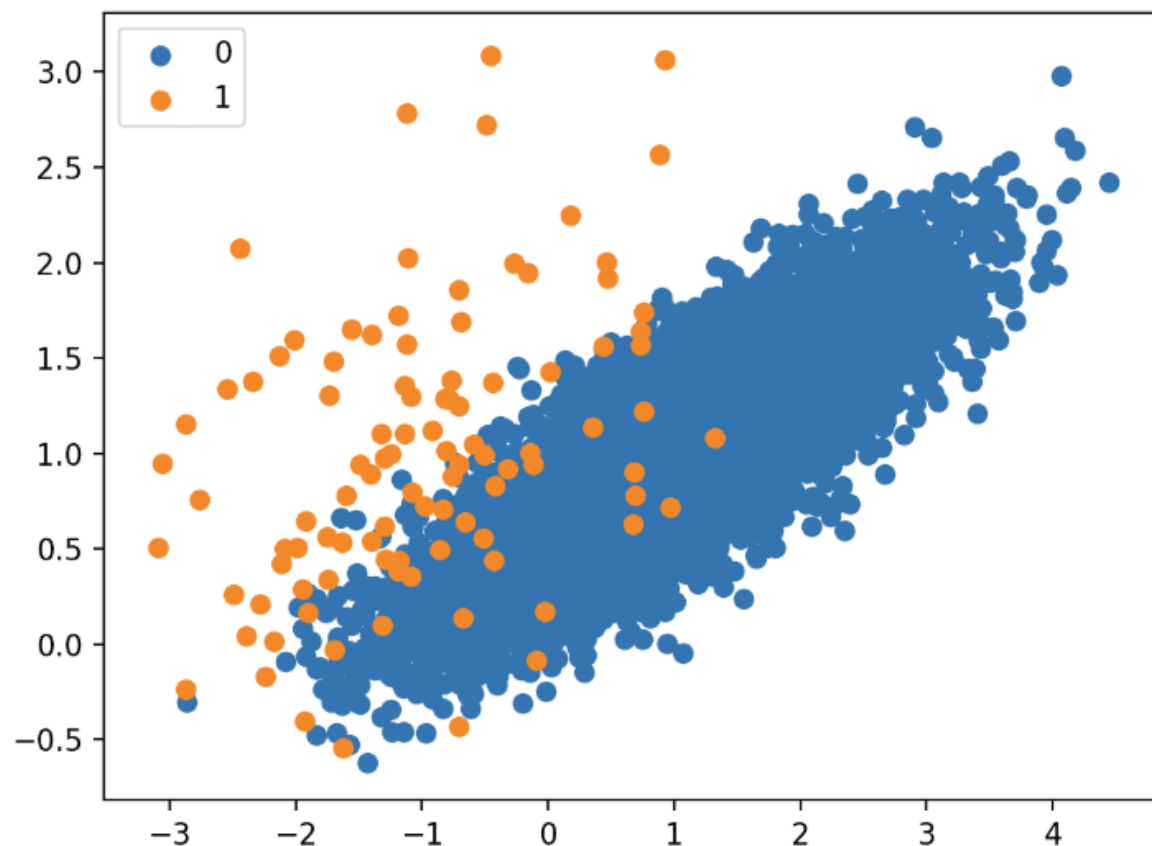
Можно перевзвешивать лосс функцию (с помощью `class_weight`)

$$\log loss = \frac{1}{N} \sum_{i=1}^N [-(w_0(y_i * \log(\hat{y}_i)) + w_1((1 - y_i) * \log(1 - \hat{y}_i)))]$$

Кроме того, хорошей практикой является подбор правильного порога



Борьба с дисбалансом. SMOTE

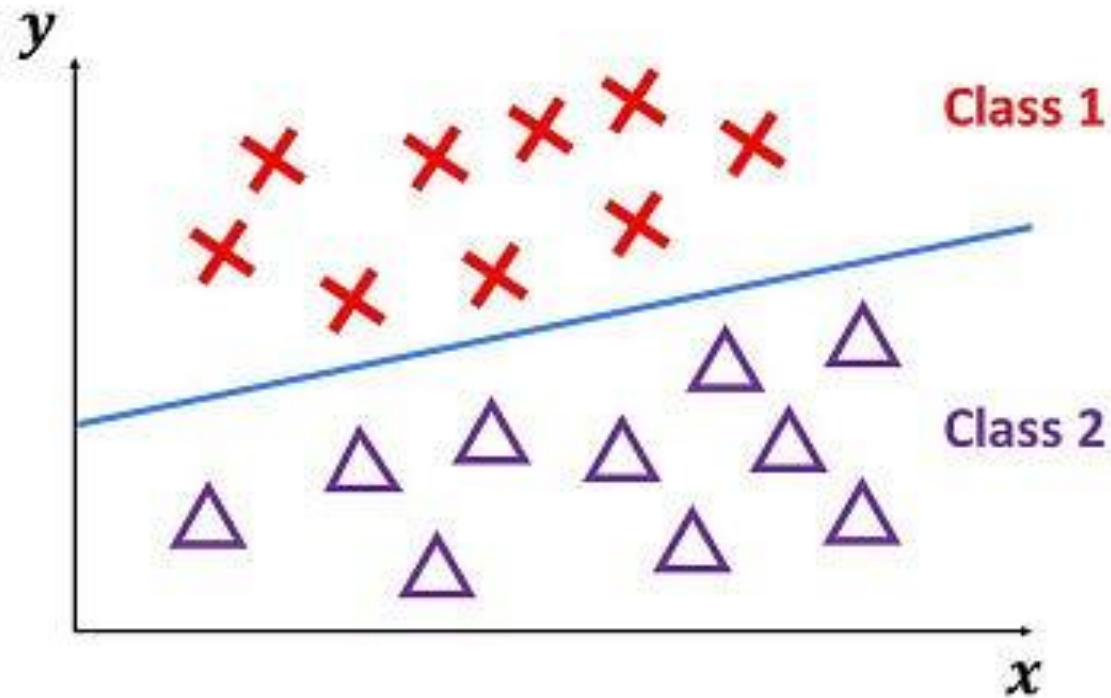


Борьба с дисбалансом. Кратко

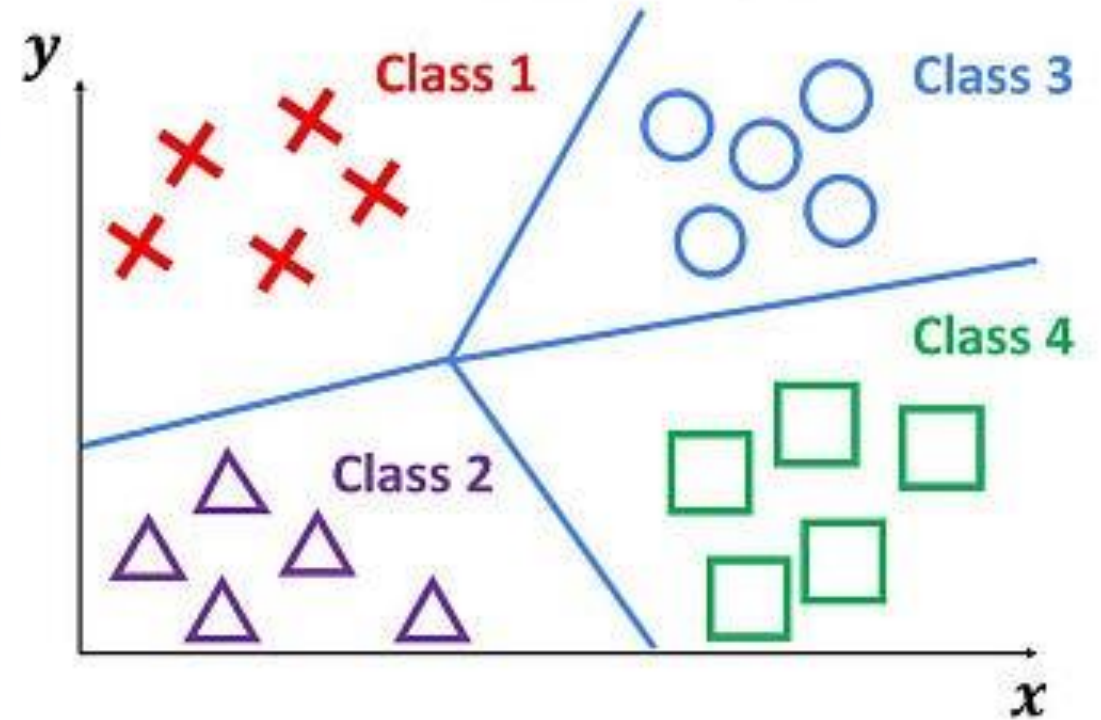
Подбор порога (без совмещения с любой другой техникой) – идеальная стратегия для «нешумных данных». На самом деле, только это и надо использовать, когда геометрия данных относительно проста, модель хорошо описывает данные (и особенно, если хорошо откалибрована). Обратим внимание, что качество признакового пространства (шум и геометрия) в классическом ML зависит исключительно от Вас, поэтому, **если Вы умеете решать задачи, то кроме подбора порога Вам ничего не нужно.**

Многоклассовая классификация

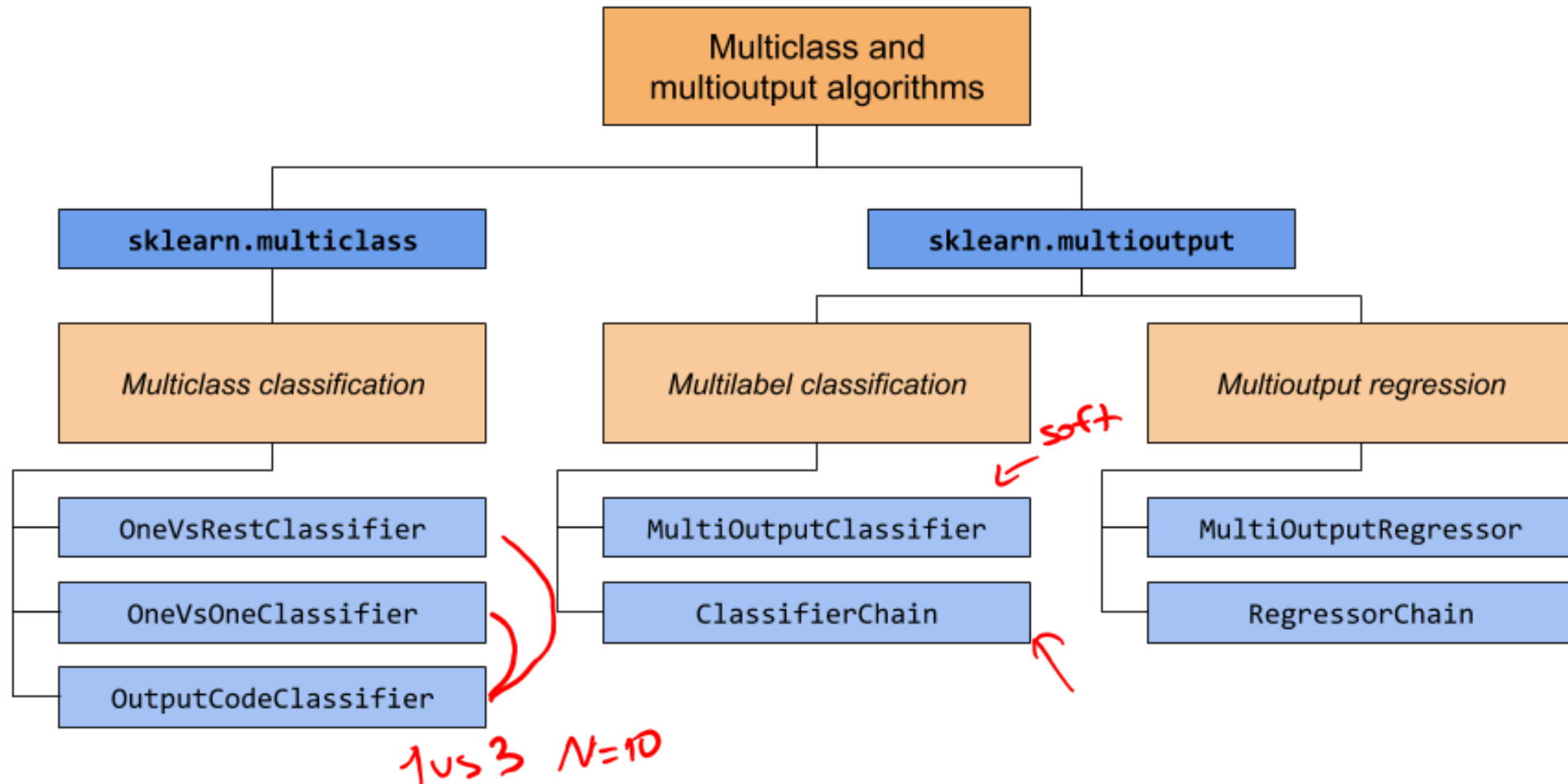
Binary Classification



Multiclass Classification












Многоклассовая классификация. Методы решения



Многоклассовая классификация

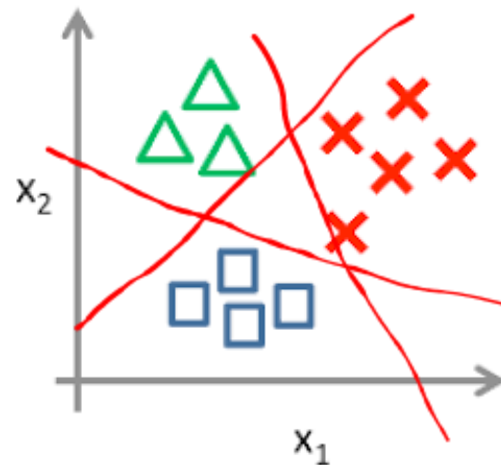
Multi-Class

Multi-Label

C = 3	Samples	Samples
  	<div data-bbox="649 692 835 871"></div> <div data-bbox="879 692 1065 871"></div> <div data-bbox="1108 692 1286 871"></div> <div data-bbox="555 942 815 1006">Labels (t)</div> <div data-bbox="657 1078 815 1135">[0 0 1]</div> <div data-bbox="879 1078 1039 1135">[1 0 0]</div> <div data-bbox="1108 1078 1268 1135">[0 1 0]</div>	<div data-bbox="1541 692 1727 871"></div> <div data-bbox="1770 692 1956 871"></div> <div data-bbox="2000 692 2186 871"></div> <div data-bbox="1447 942 1707 1006">Labels (t)</div> <div data-bbox="1559 1078 1719 1135">[1 0 1]</div> <div data-bbox="1783 1078 1944 1135">[0 1 0]</div> <div data-bbox="2007 1078 2168 1135">[1 1 1]</div>

Многоклассовая классификация. Multiclass

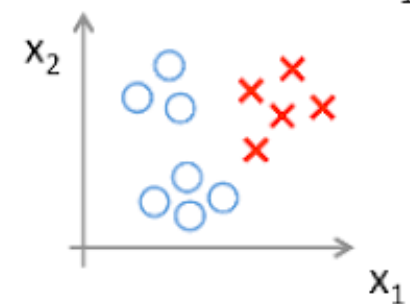
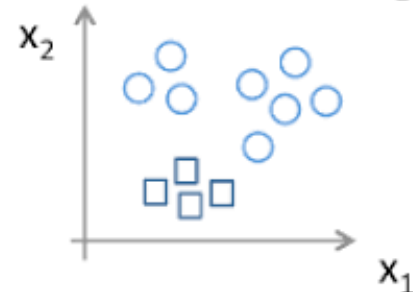
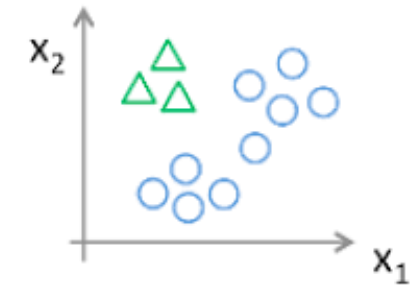
One-vs-all (one-vs-rest):



Class 1: **Green**

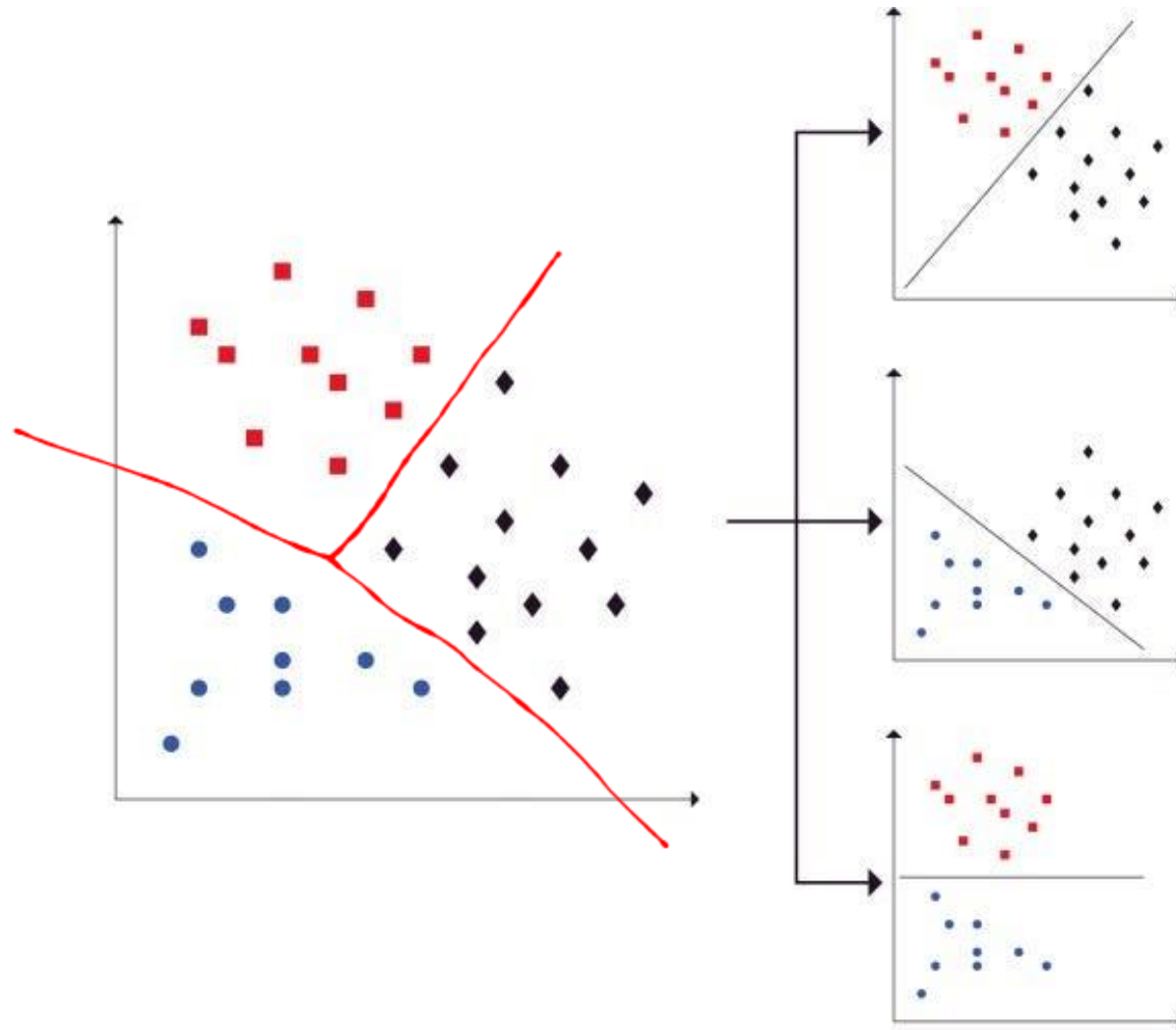
Class 2: **Blue**

Class 3: **Red**

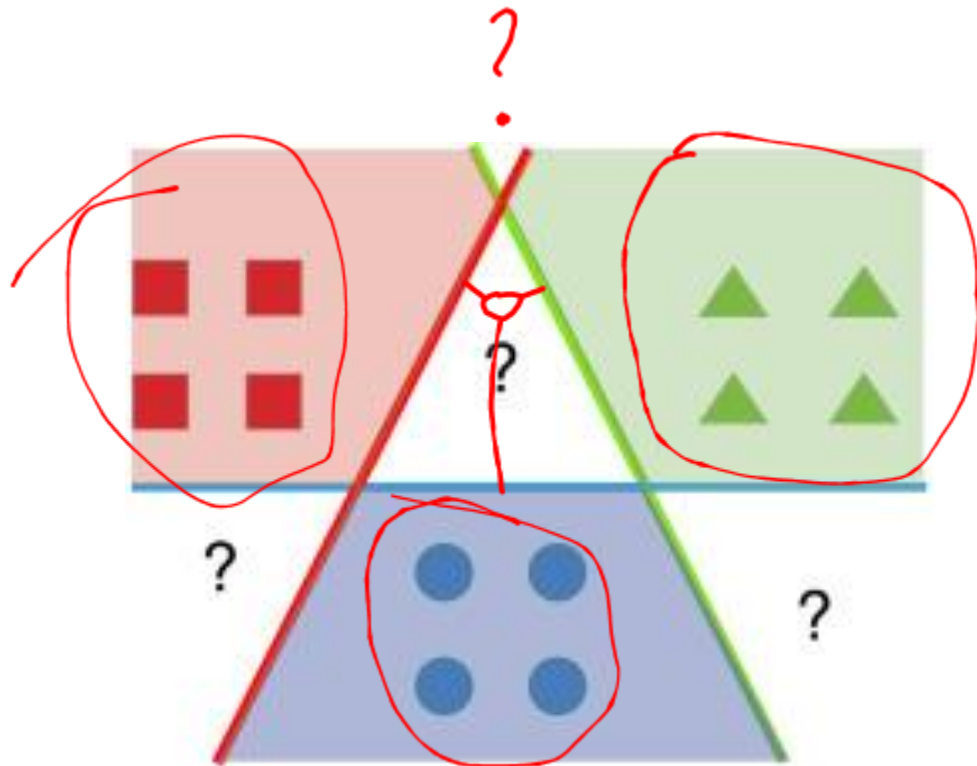


Многоклассовая классификация. Multiclass

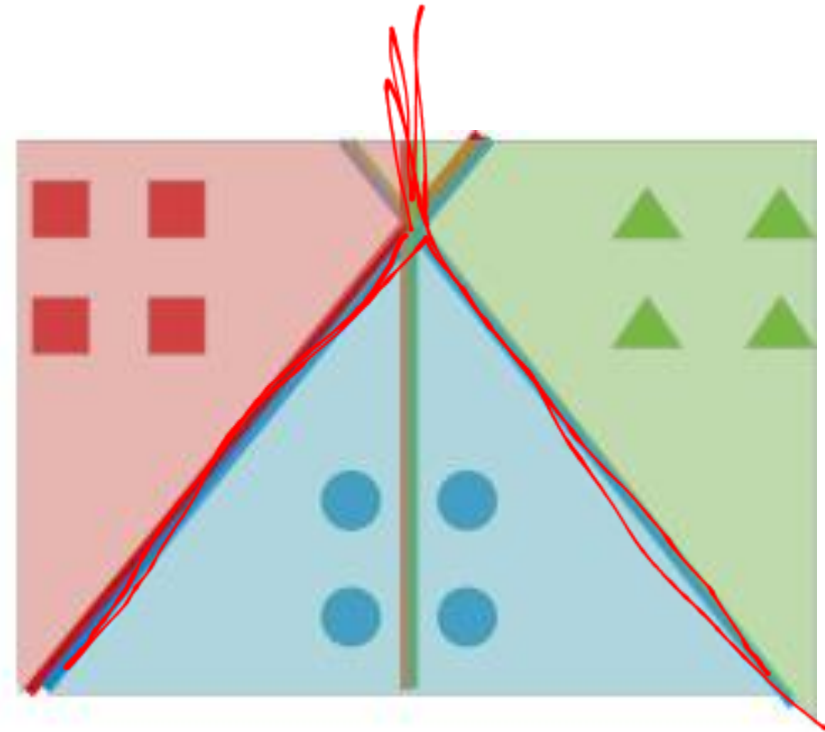
One VS One



OvA & OvO



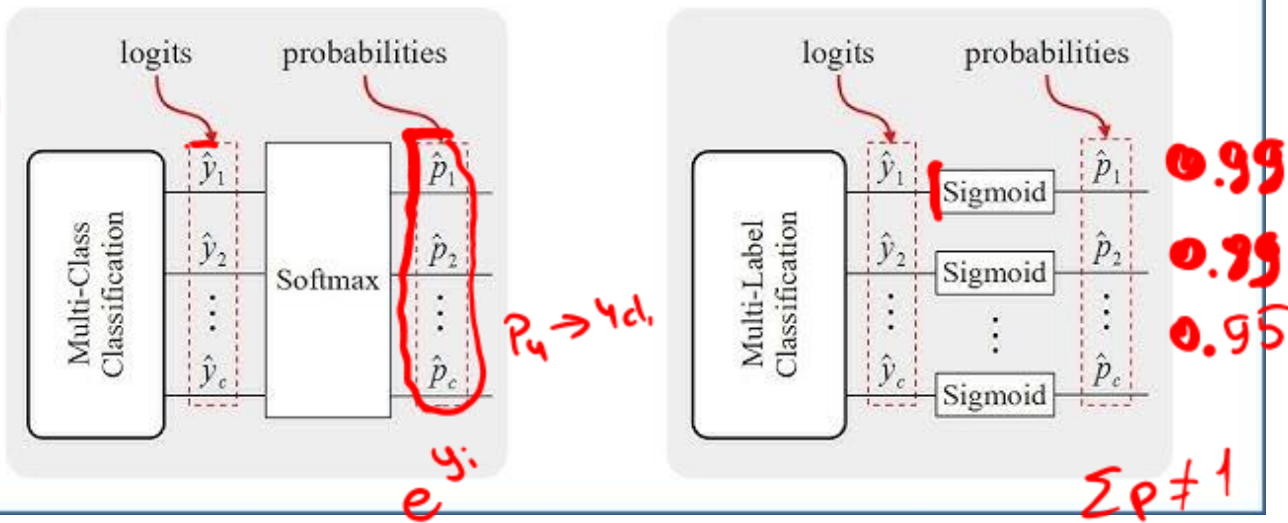
(a) Separation with OvA.



(b) Separation with OvO.

Многоклассовая классификация. Multilabel

MULTI-CLASS vs. MULTI-LABEL CLASSIFICATION



X	Y1	Y2	Y3	Y4
x(1)	0	1	0	1
x(2)	0	0	1	1
x(3)	1	1	1	0

Data Set

X	Y1
x(1)	0
x(2)	0
x(3)	1

Classifier 1

X	Y1	Y2	Y3
x(1)	0	1	0
x(2)	0	0	1
x(3)	1	1	1

Classifier 3

X	Y1	Y2
x(1)	0	1
x(2)	0	0
x(3)	1	1

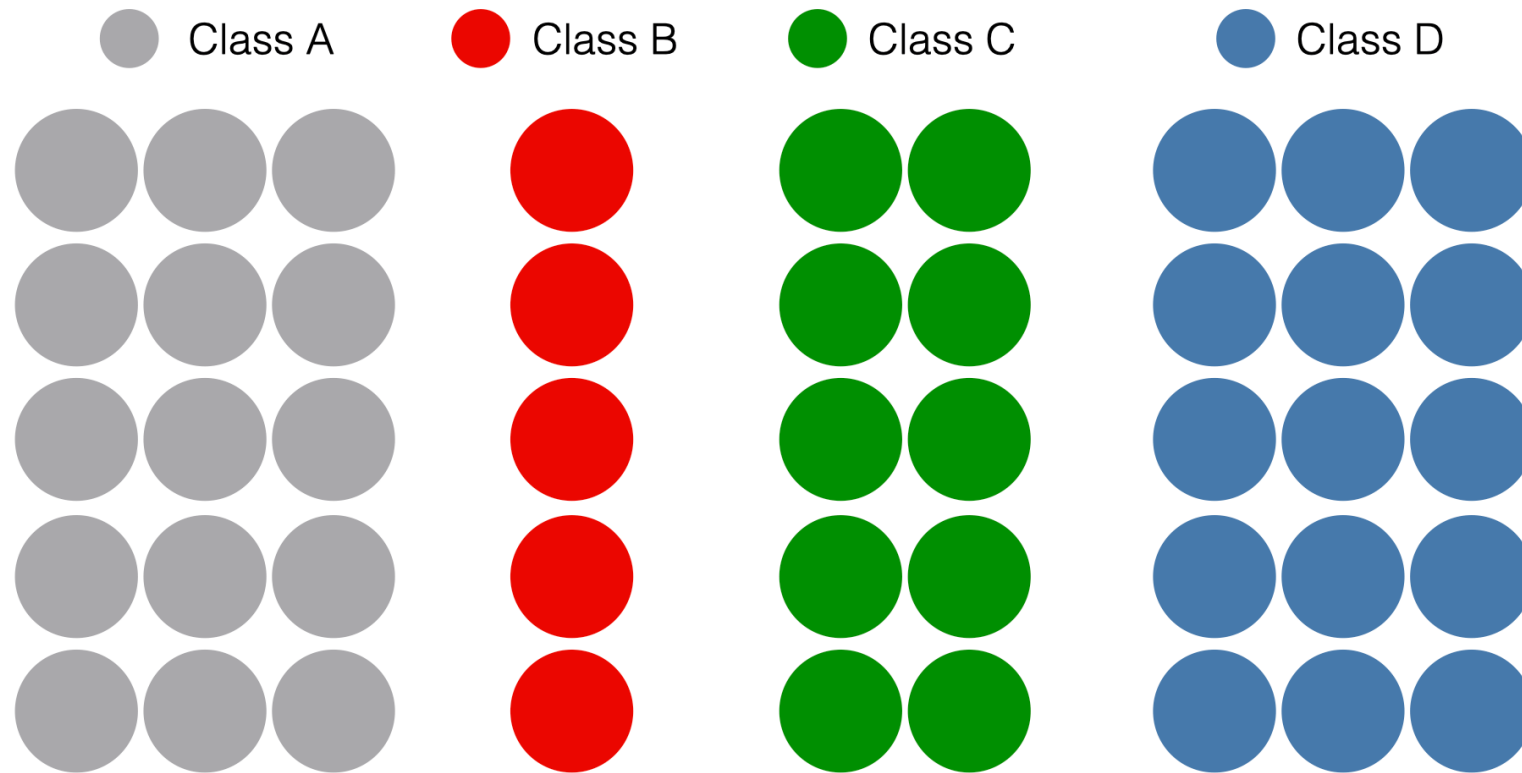
Classifier 2

X	Y1	Y2	Y3	Y4
x(1)	0	1	0	1
x(2)	0	0	1	1
x(3)	1	1	1	0

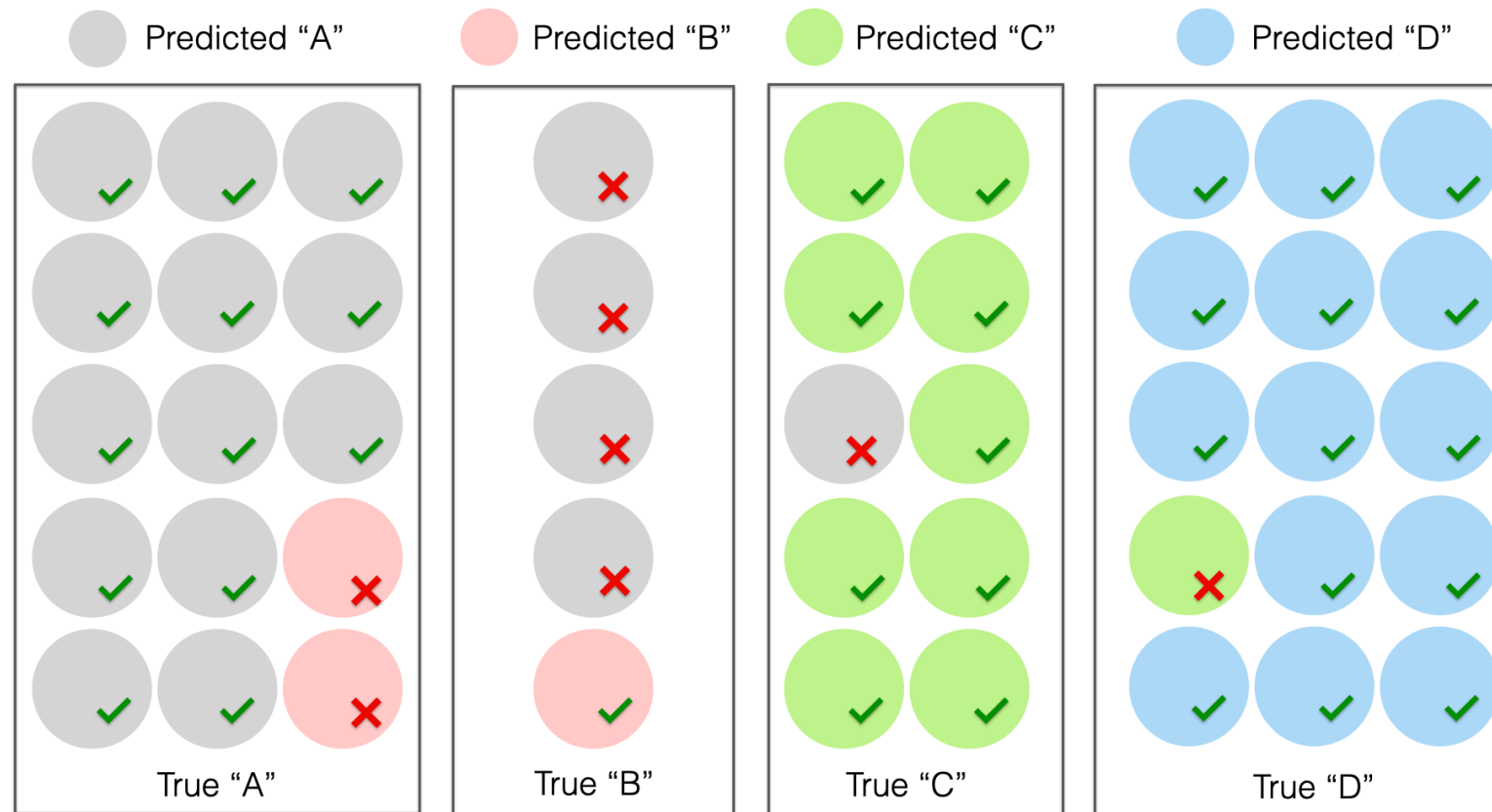
Classifier 4

cat

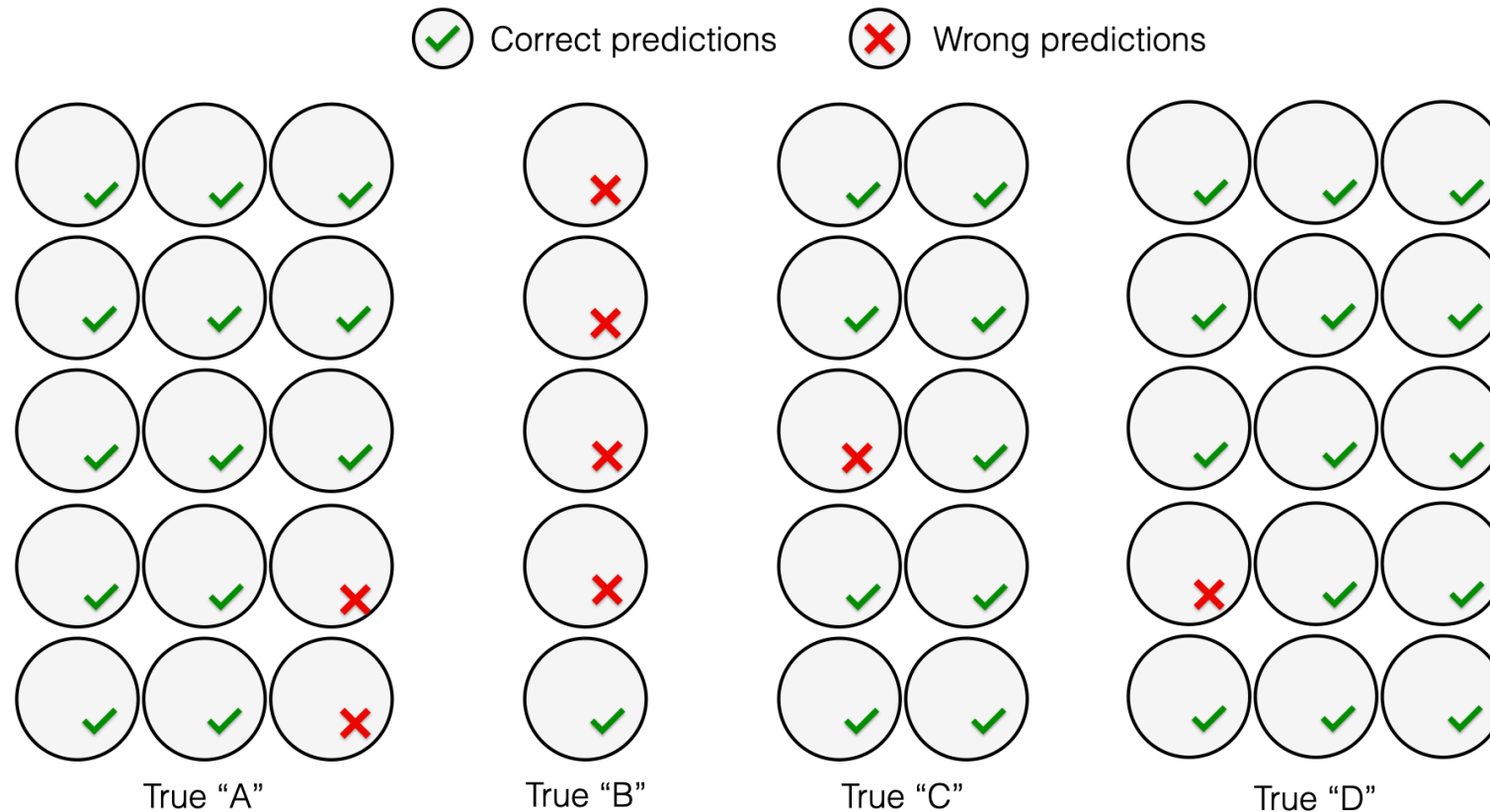
Многоклассовая классификация. Метрики качества



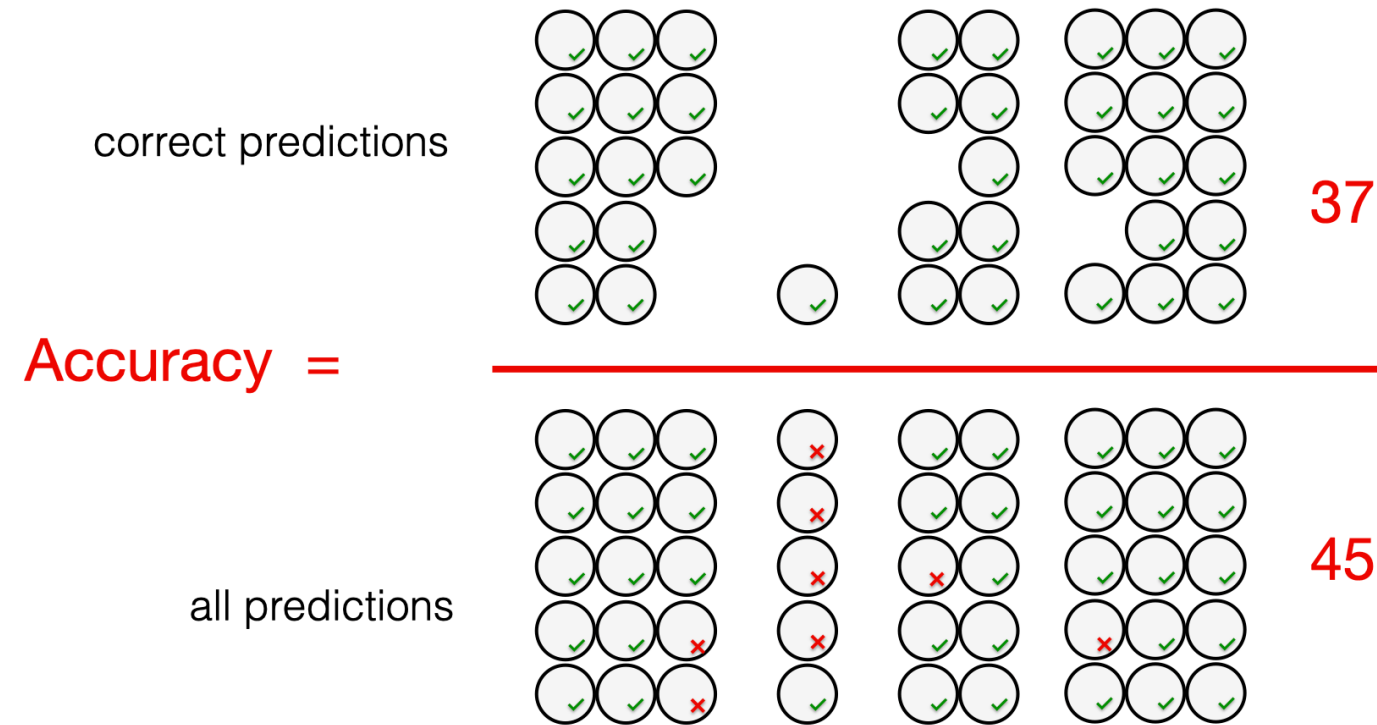
Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества

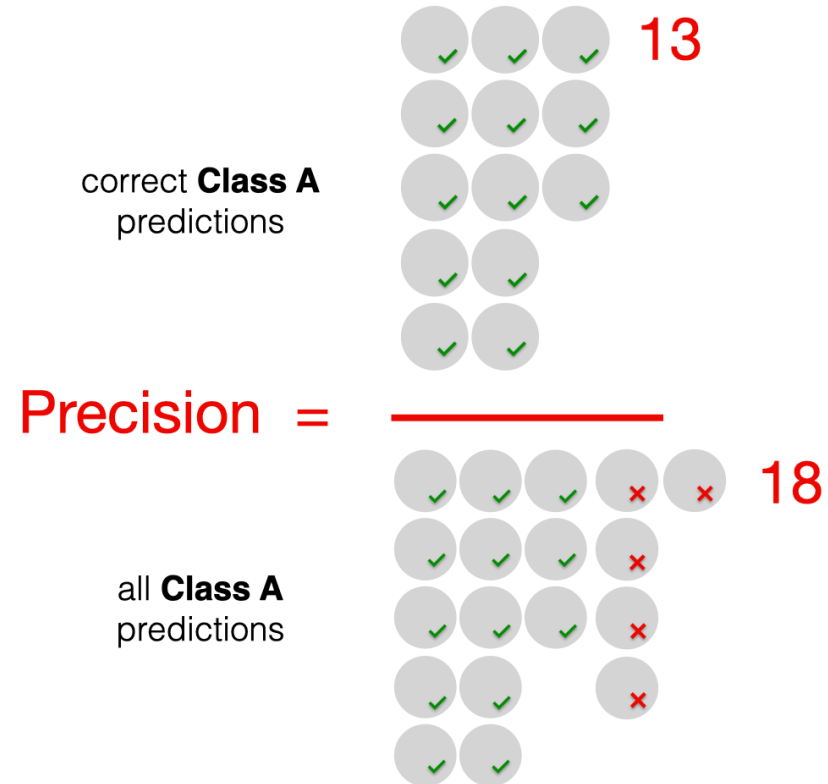
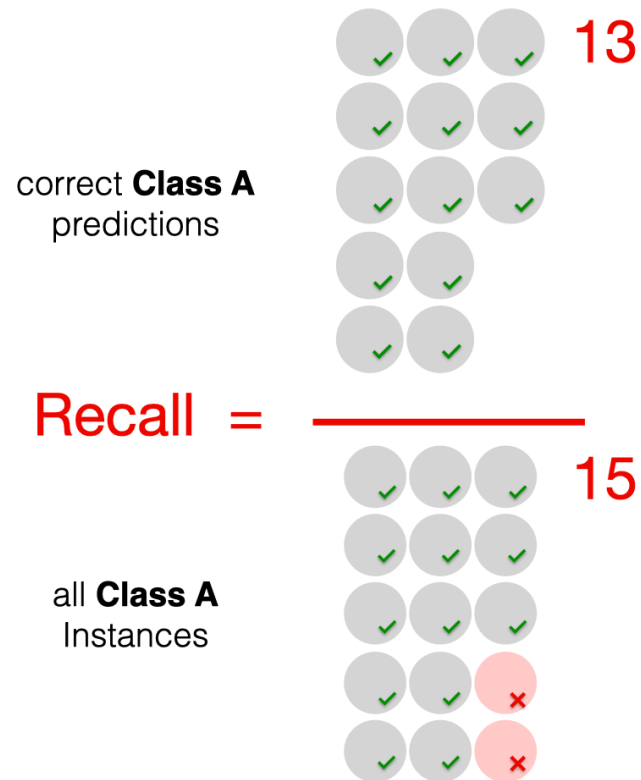
$$\text{Recall}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FN_{\text{Class A}}}$$



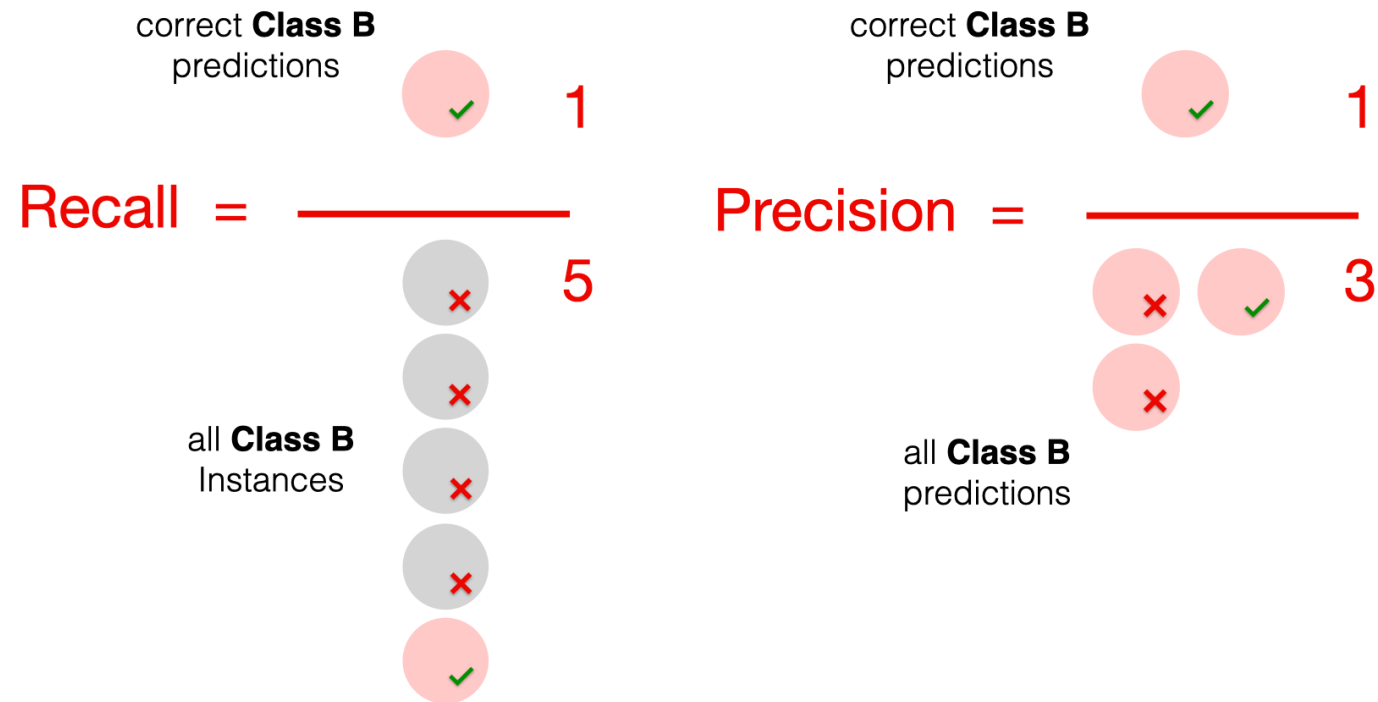
$$\text{Precision}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FP_{\text{Class A}}}$$



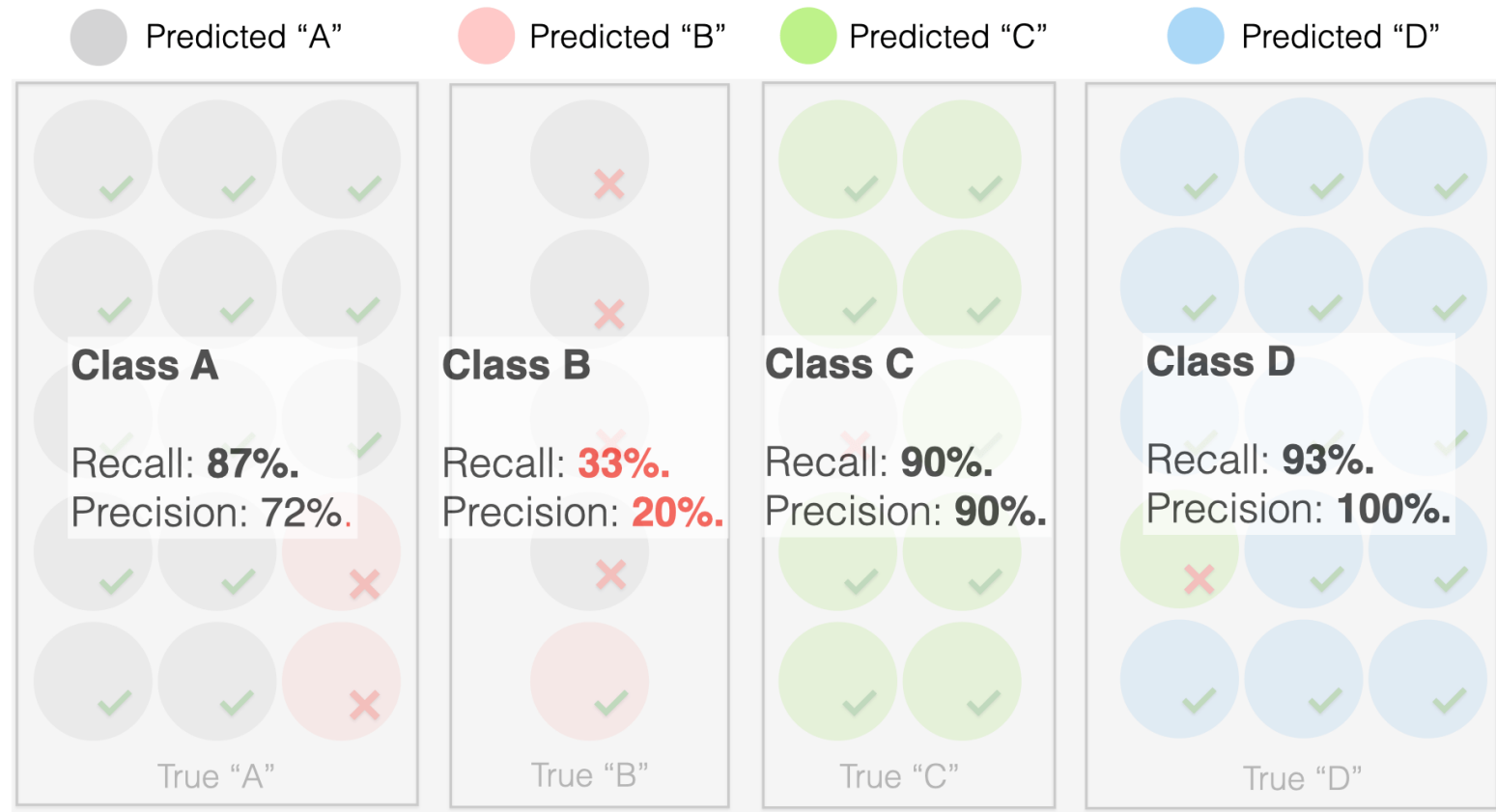
Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества



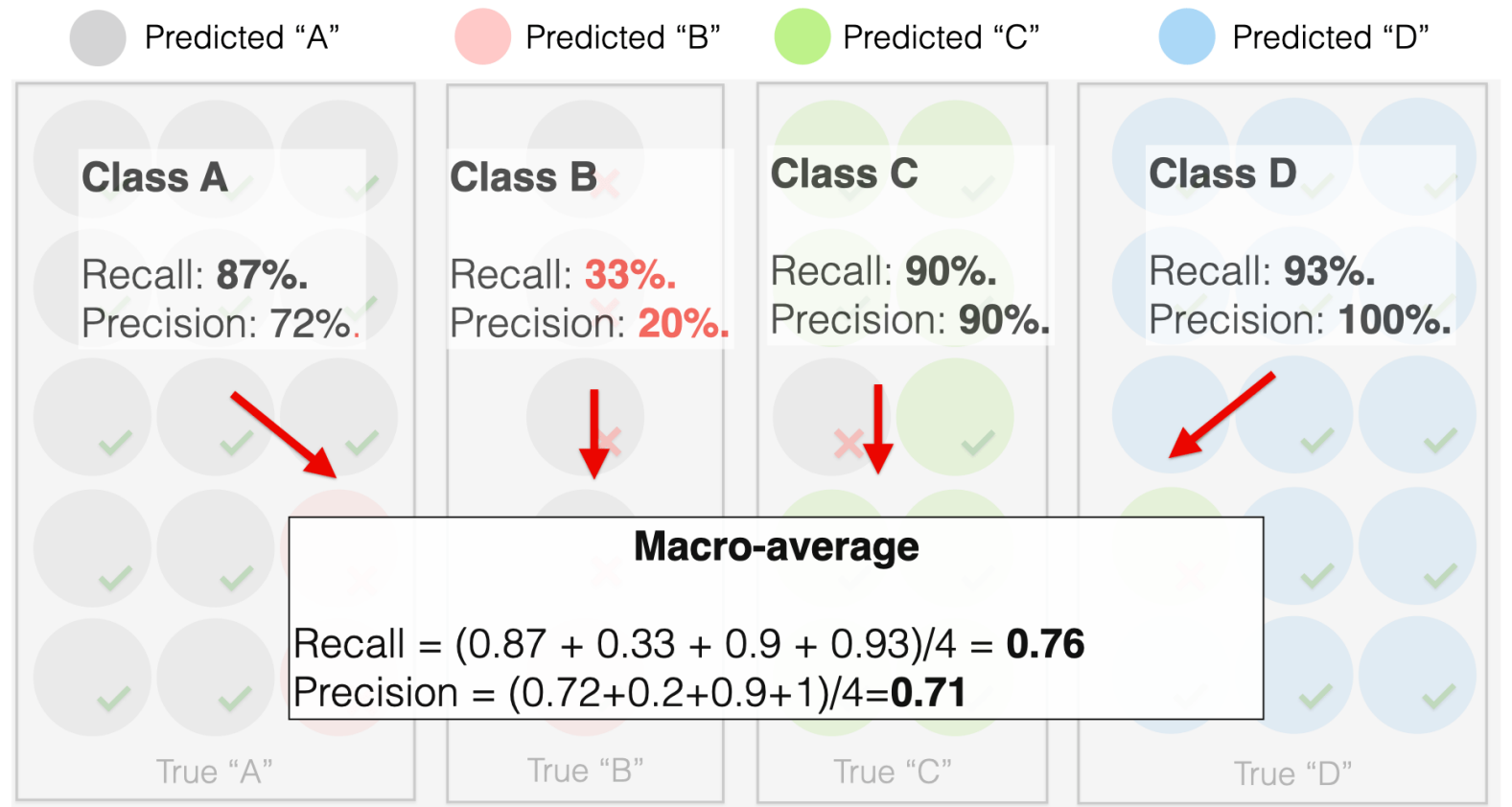
Многоклассовая классификация. Метрики качества

$$\text{Precision}_{\text{Macro-average}} = \frac{\text{Precision}_{\text{Class A}} + \text{Precision}_{\text{Class B}} + \dots + \text{Precision}_{\text{Class N}}}{N}$$

$$\text{Recall}_{\text{Macro-average}} = \frac{\text{Recall}_{\text{Class A}} + \text{Recall}_{\text{Class B}} + \dots + \text{Recall}_{\text{Class N}}}{N}$$



Многоклассовая классификация. Метрики качества



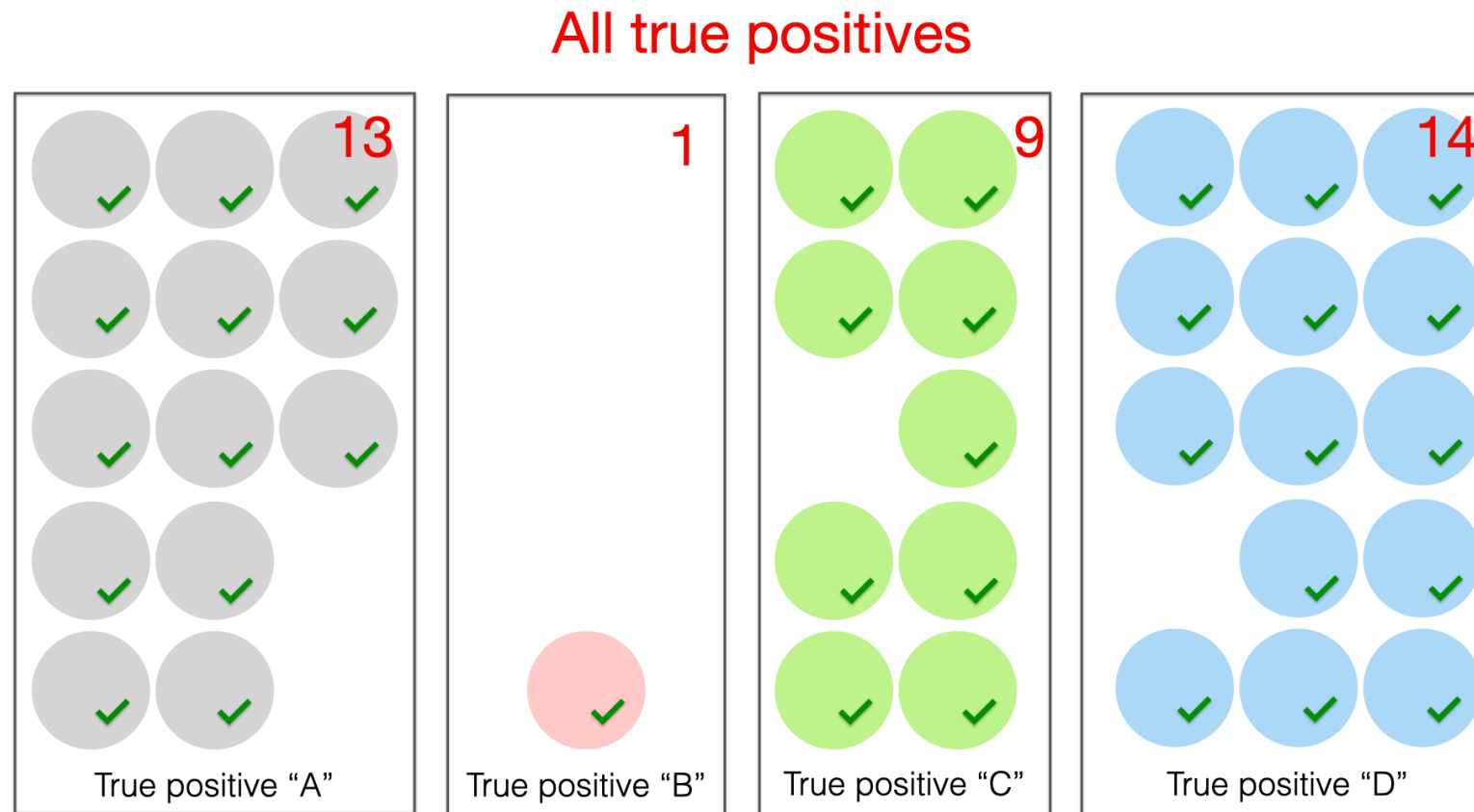
Многоклассовая классификация. Метрики качества

$$\text{Precision}_{\text{Micro-average}} = \frac{TP_A + TP_B + \dots TP_N}{TP_A + FP_A + TP_B + FP_B + \dots TP_N + FP_N}$$

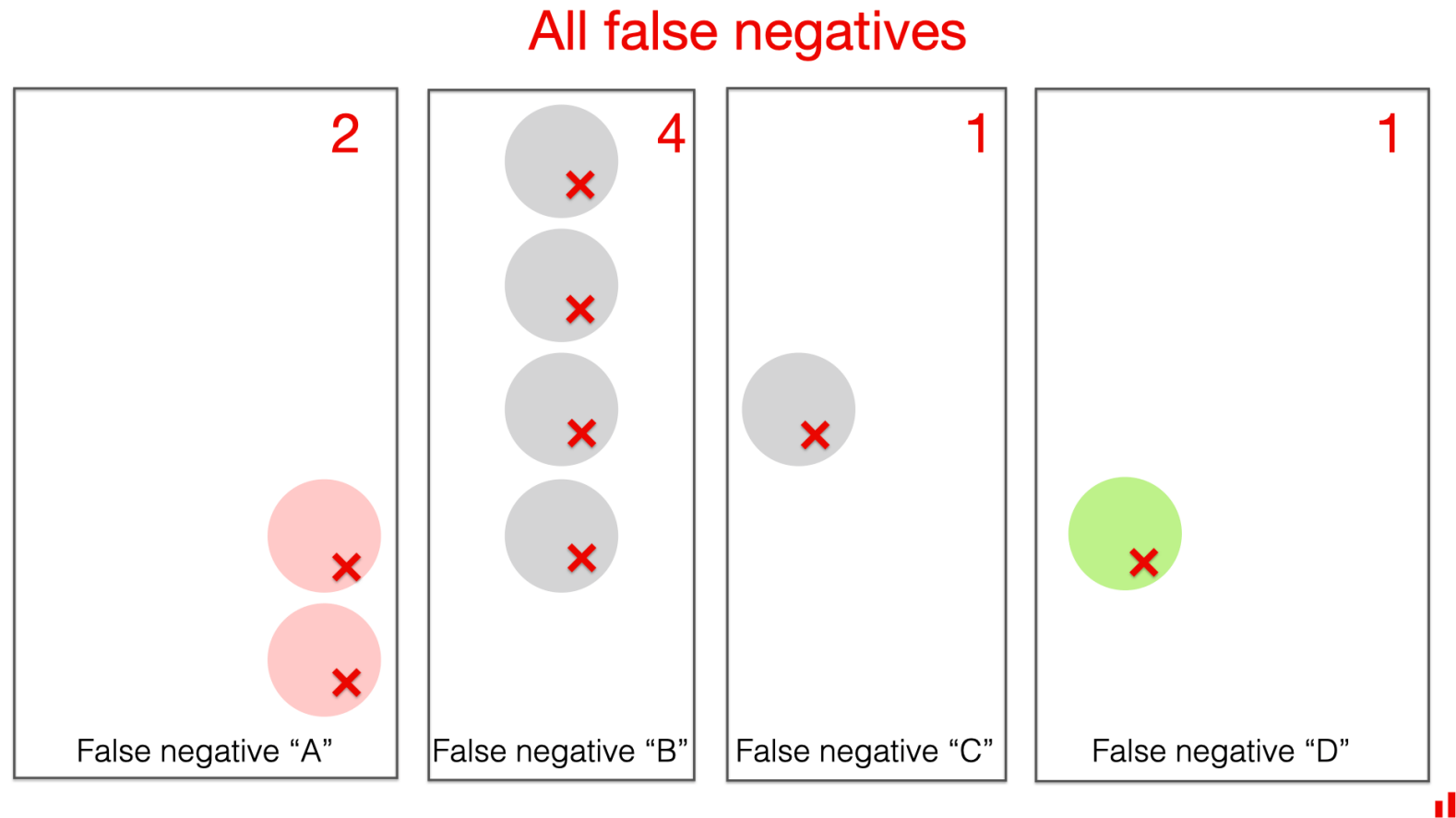
$$\text{Recall}_{\text{Micro-average}} = \frac{TP_A + TP_B + \dots TP_N}{TP_A + FN_A + TP_B + FN_B + \dots TP_N + FN_N}$$



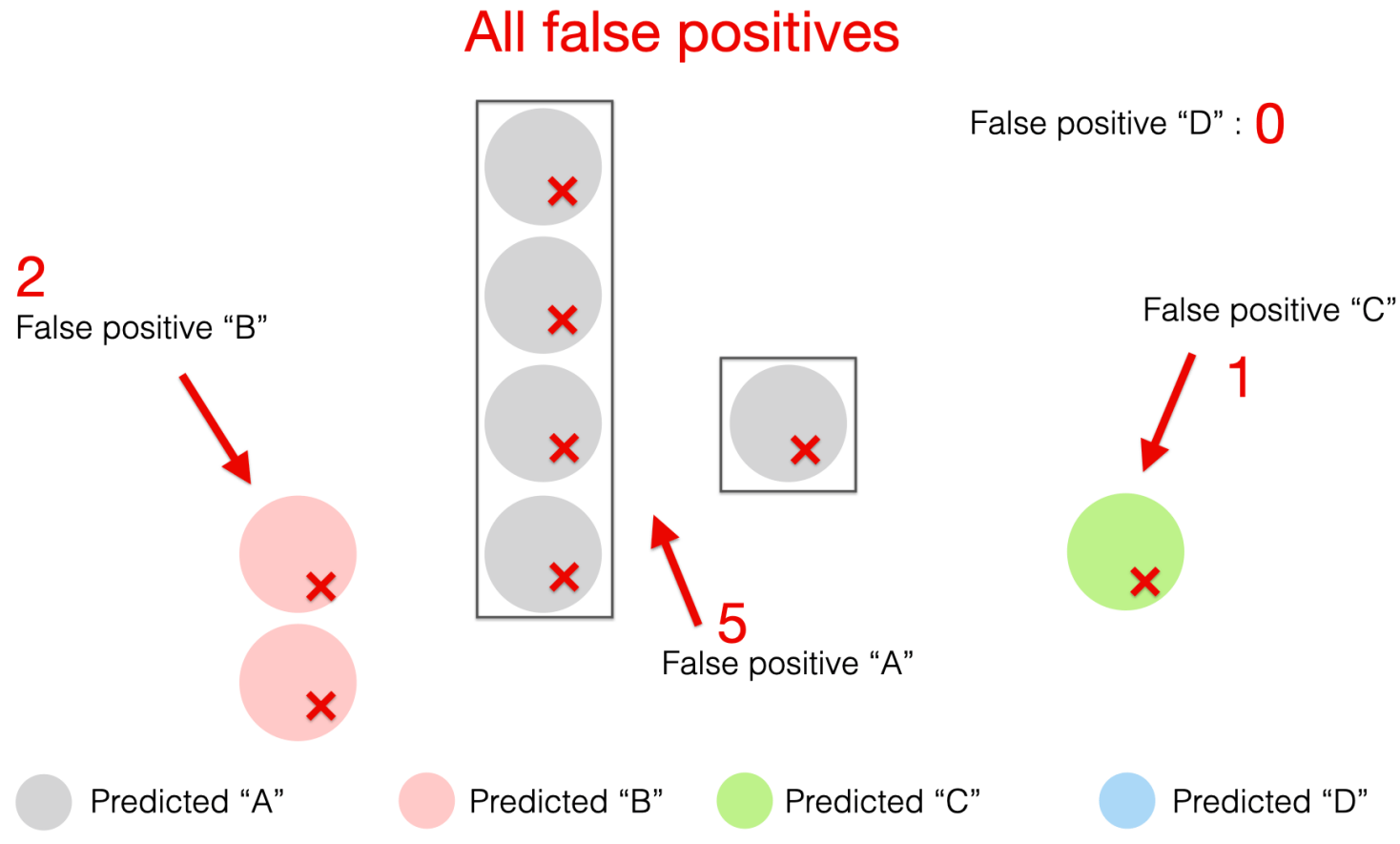
Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества



Многоклассовая классификация. Метрики качества

	Total TP	Total FP	Total FN
	13 + 1 + 9 + 14	2 + 5 + 1 + 0	2 + 4 + 1 + 1

Precision = $\frac{13 + 1 + 9 + 14}{13 + 1 + 9 + 14 + 2 + 5 + 1 + 0} = 0.82$
Micro-average

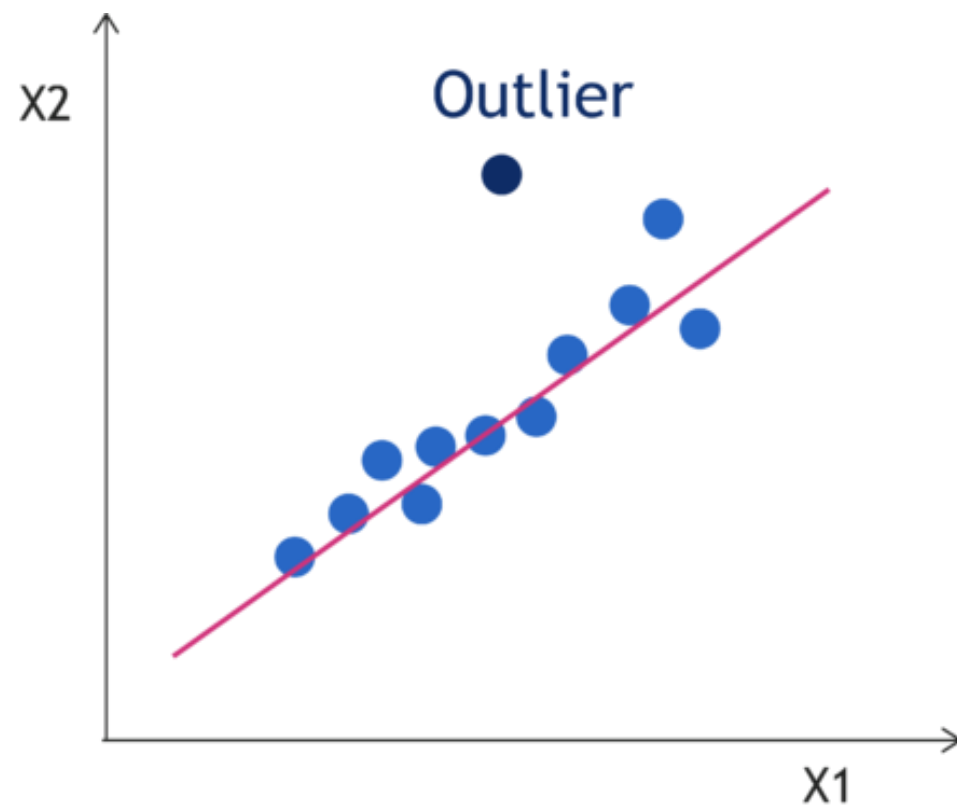
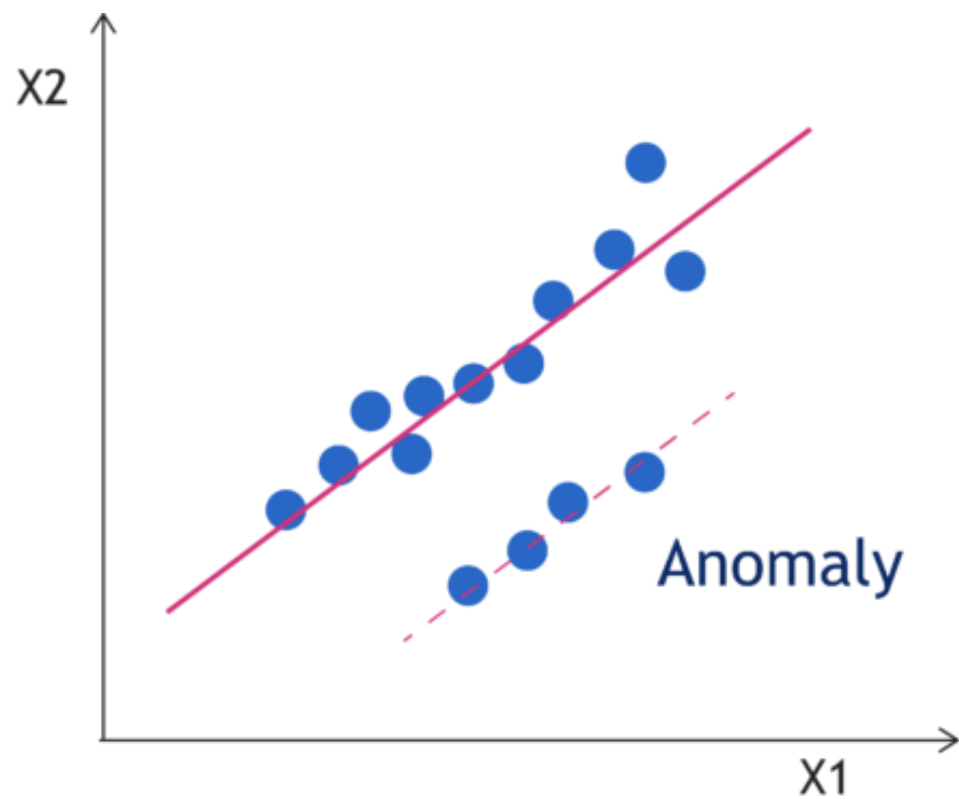
Recall = $\frac{13 + 1 + 9 + 14}{13 + 1 + 9 + 14 + 2 + 4 + 1 + 1} = 0.82$
Micro-average



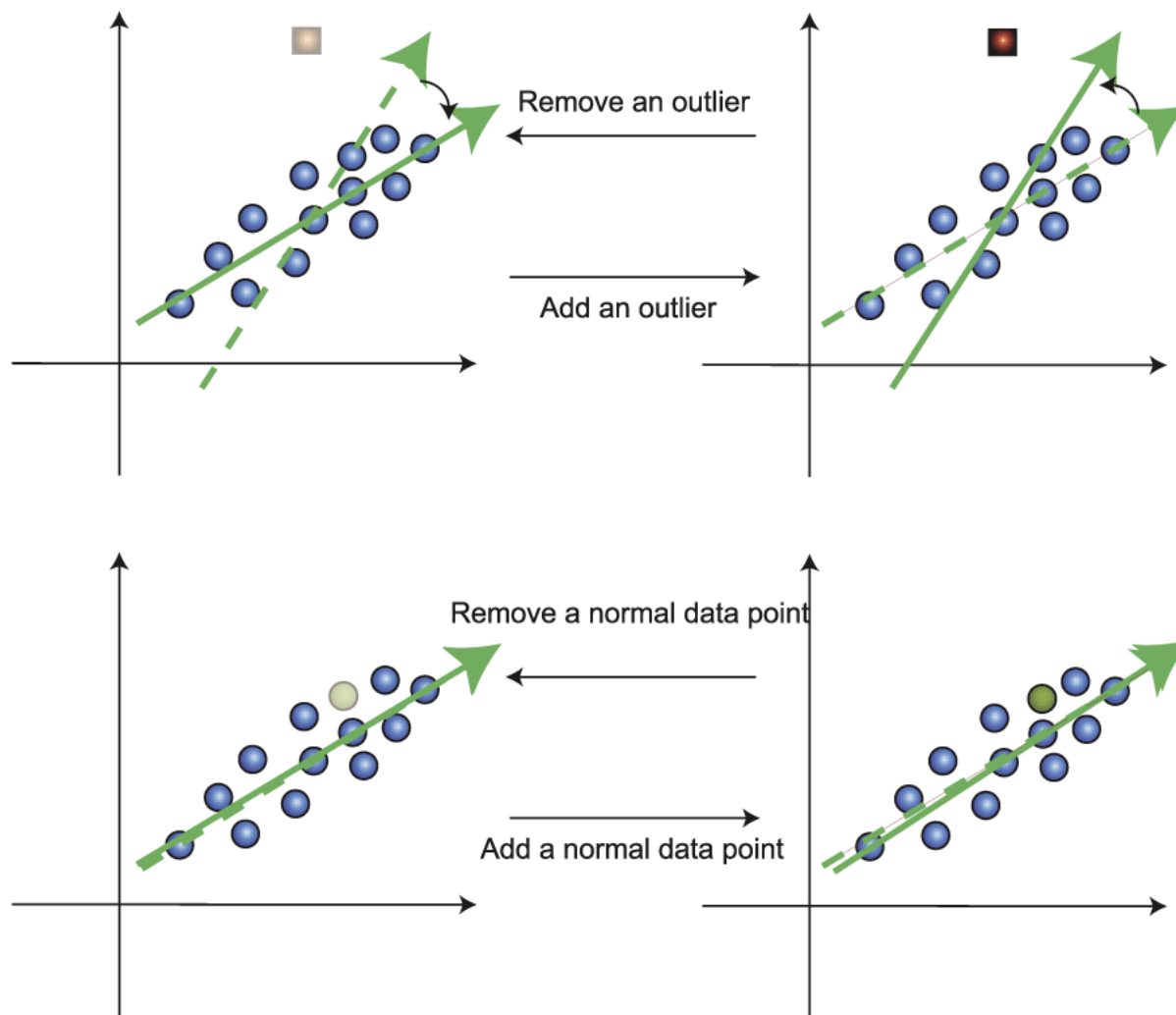
Многоклассовая классификация. Метрики качества

Критерий	Макро	Микро
Относительно чего смотрим	Классы (все равны, неважно, сколько наблюдений в каждом классе)	Объекты (все равны, неважно, какого класса каждый объект)
Полезно в случае	Все классы одинаково важны Есть дисбаланс классов	Важен общий уровень ошибок
Возможные недостатки	Влияние редких классов может сильно портить картину	Можно пропустить плохую работу на малых классах
Искажения	Заниженный результат из-за слабых классов Размывание ошибки в важных (частых) классах	Завышение результата за счет доминирующего класса

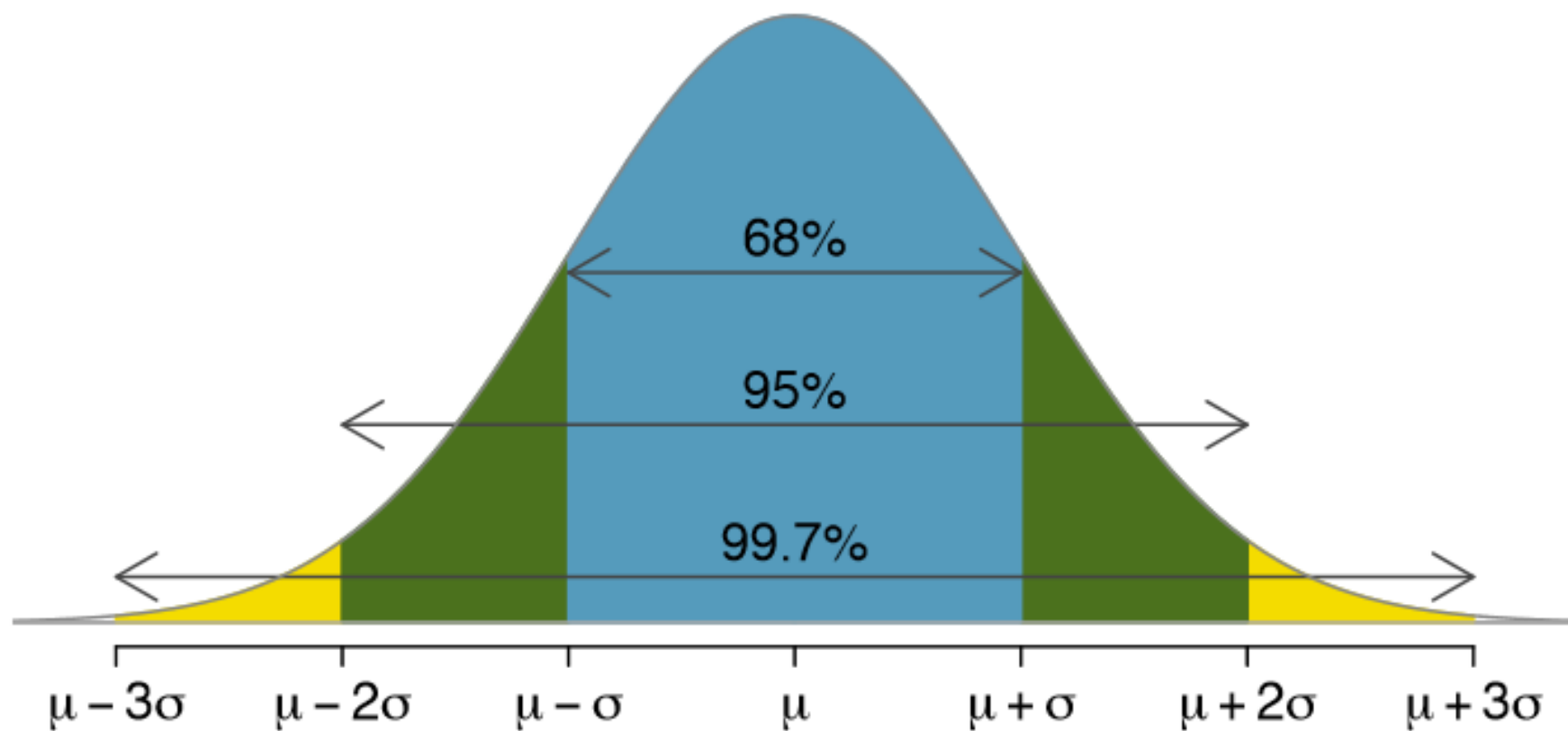
Поиск аномалий



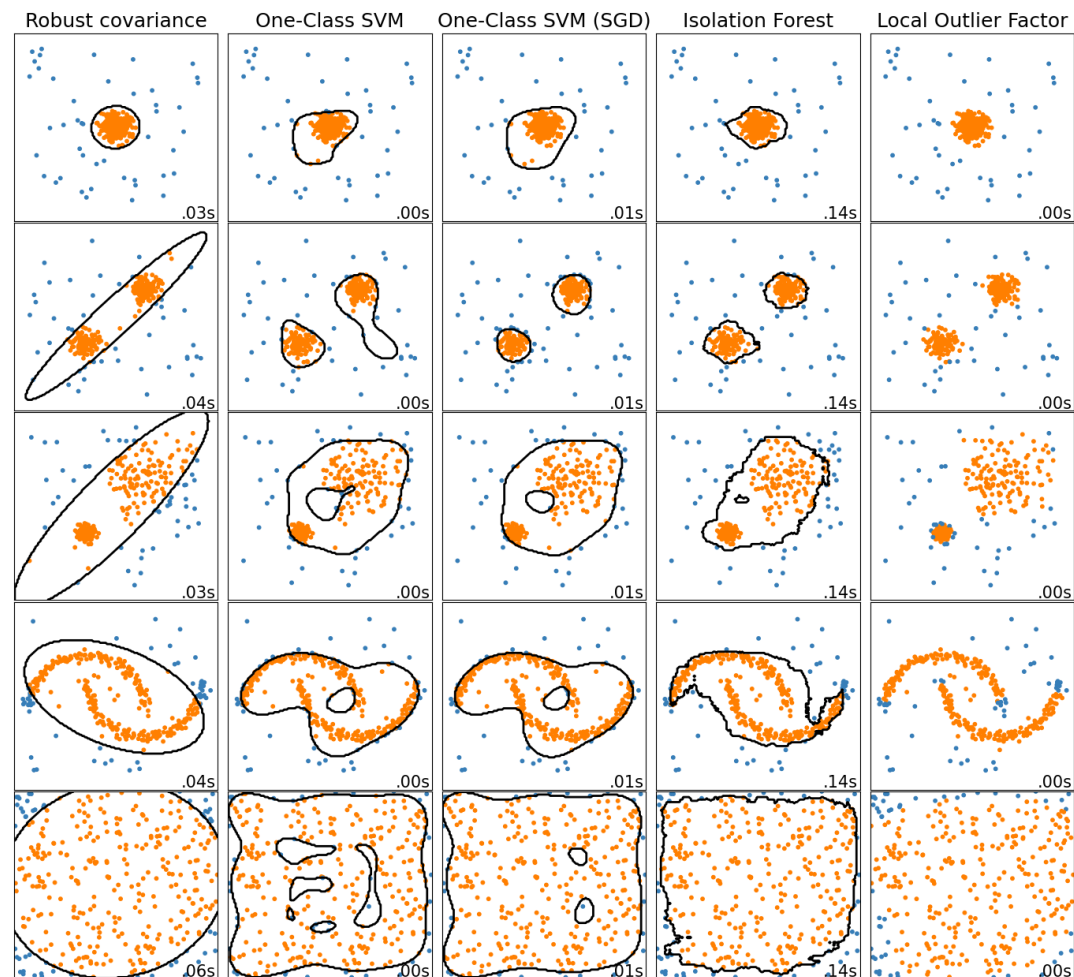
Поиск аномалий. Зачем?



Поиск аномалий. Как?



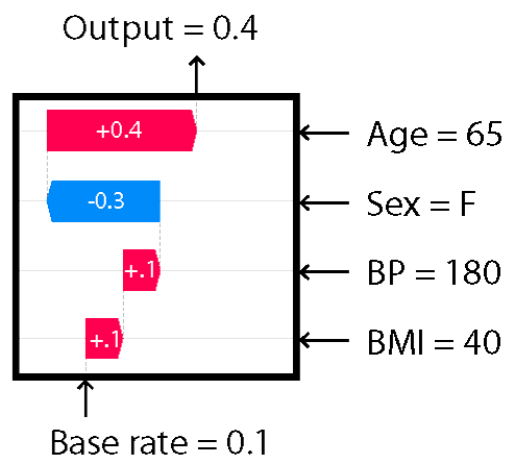
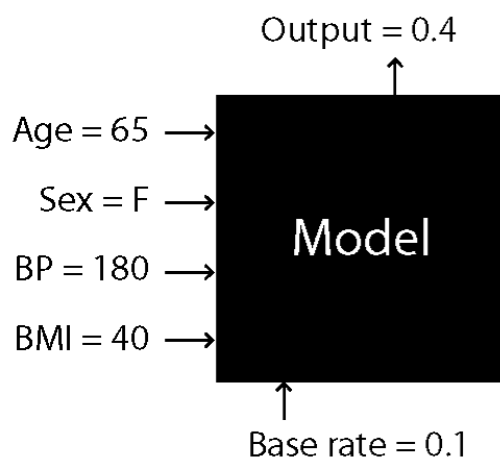
Поиск аномалий. Как?



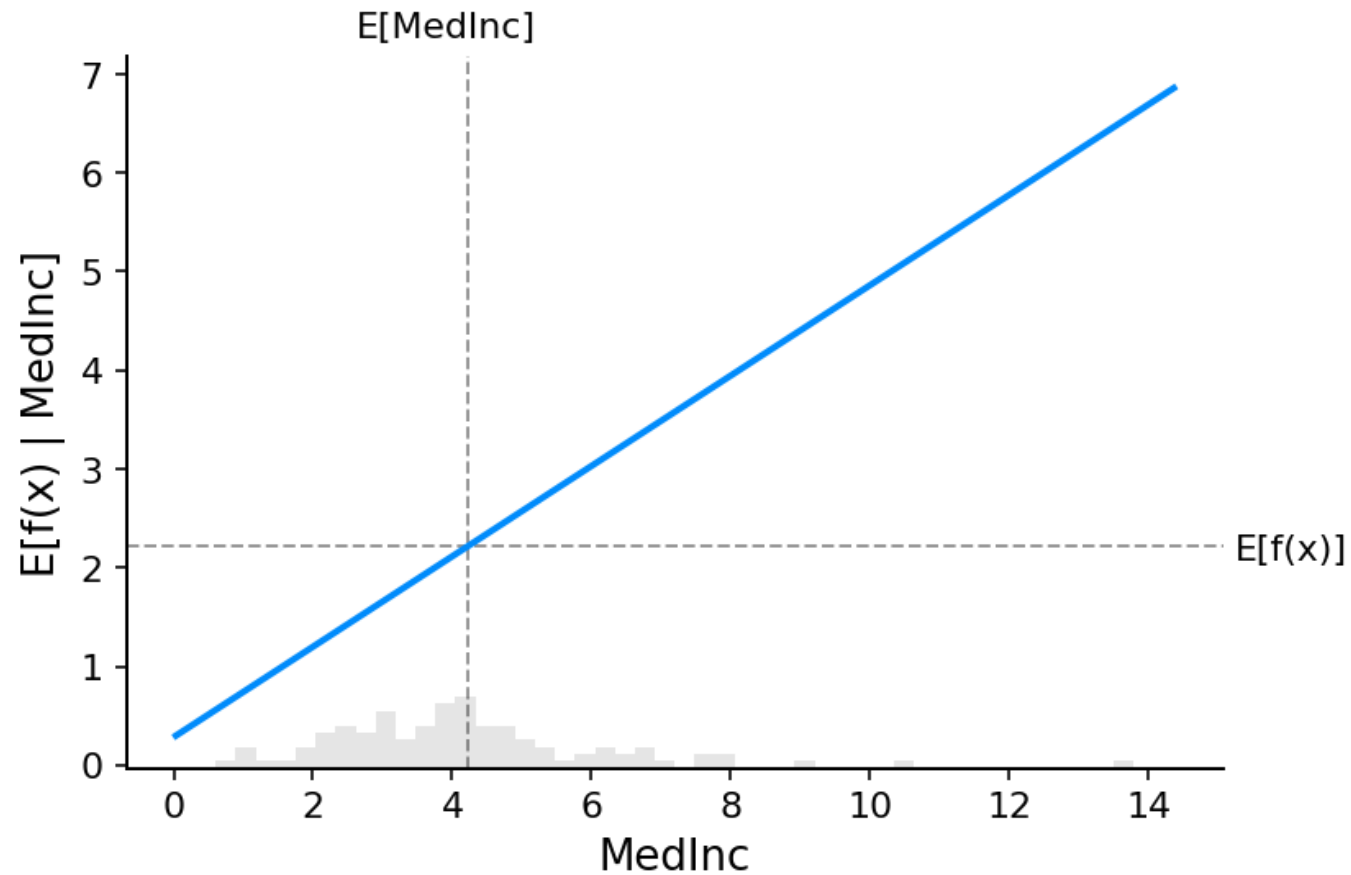
SHAP



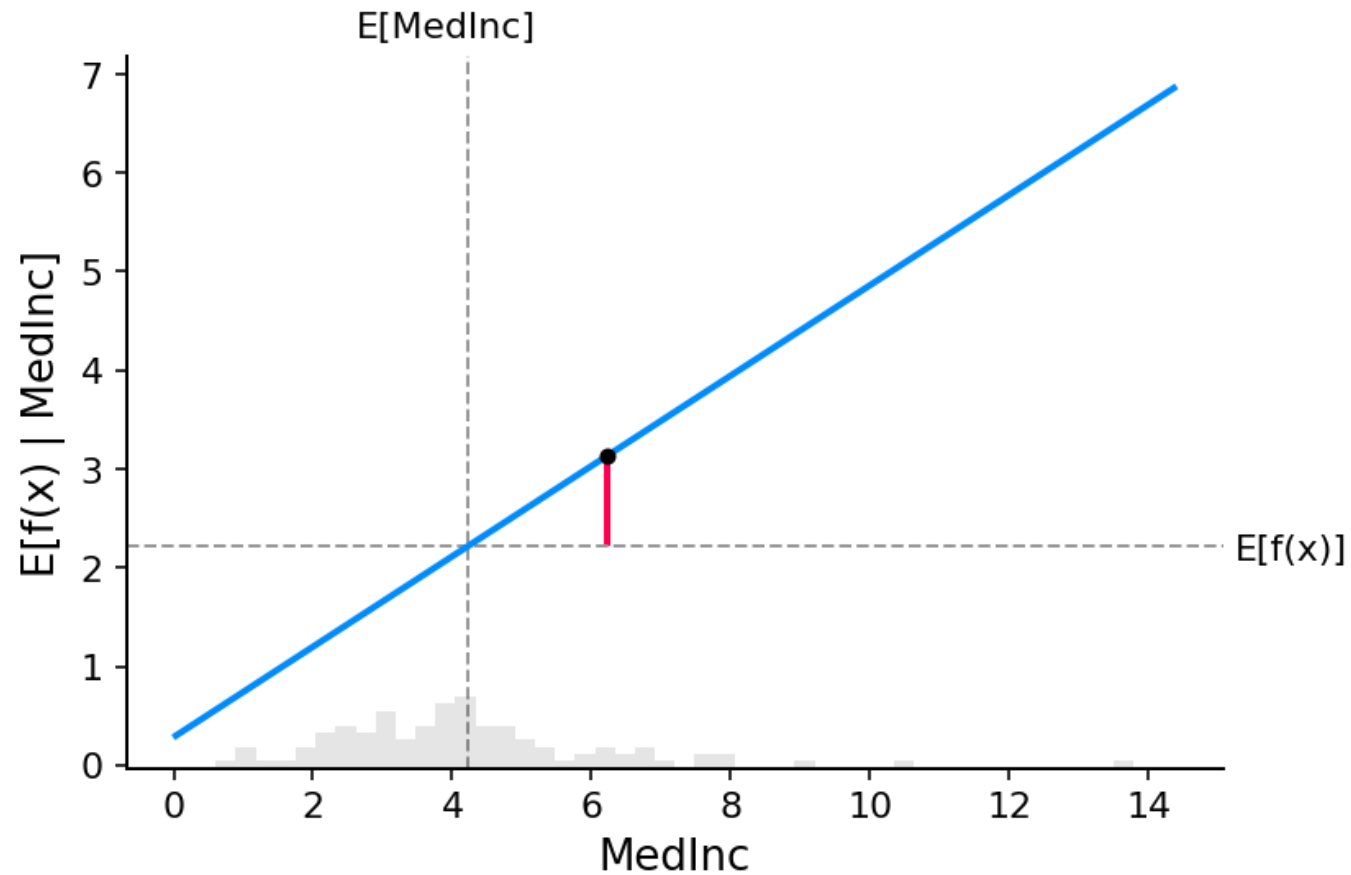
SHAP



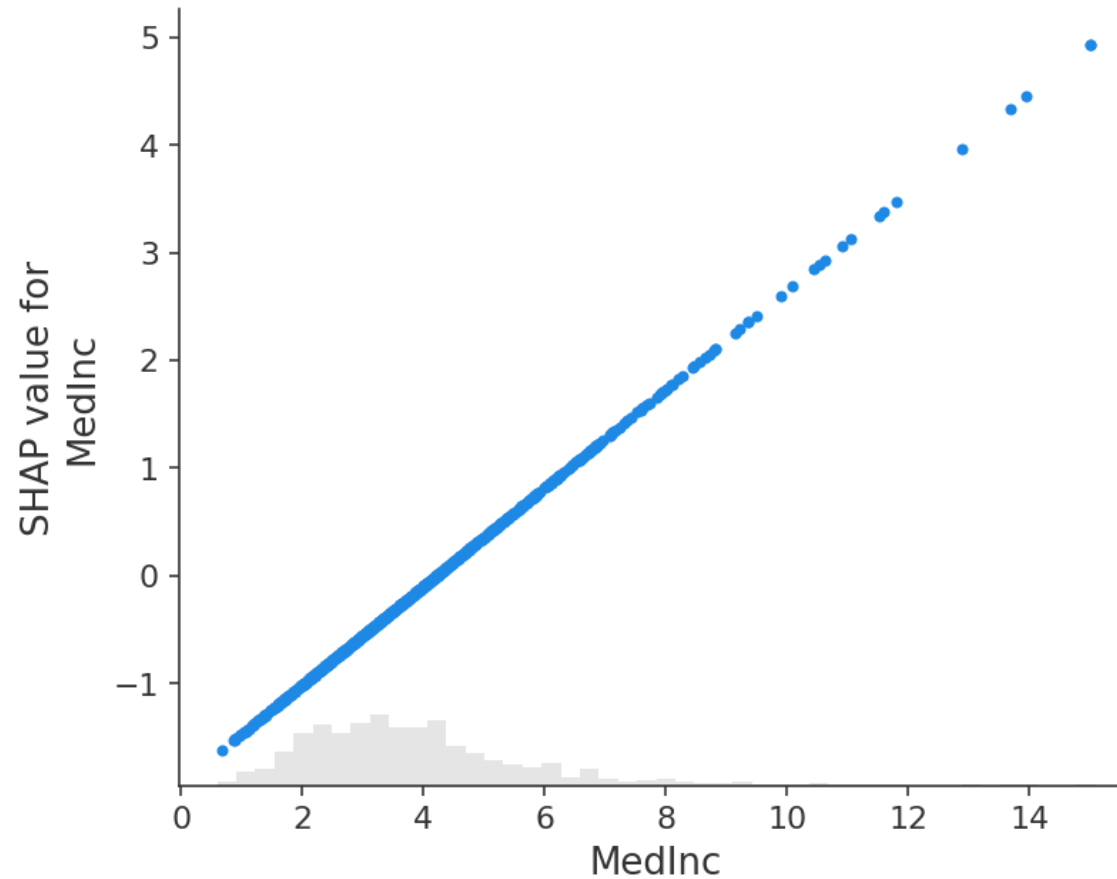
SHAP. График частичной зависимости



SHAR. График частичной зависимости

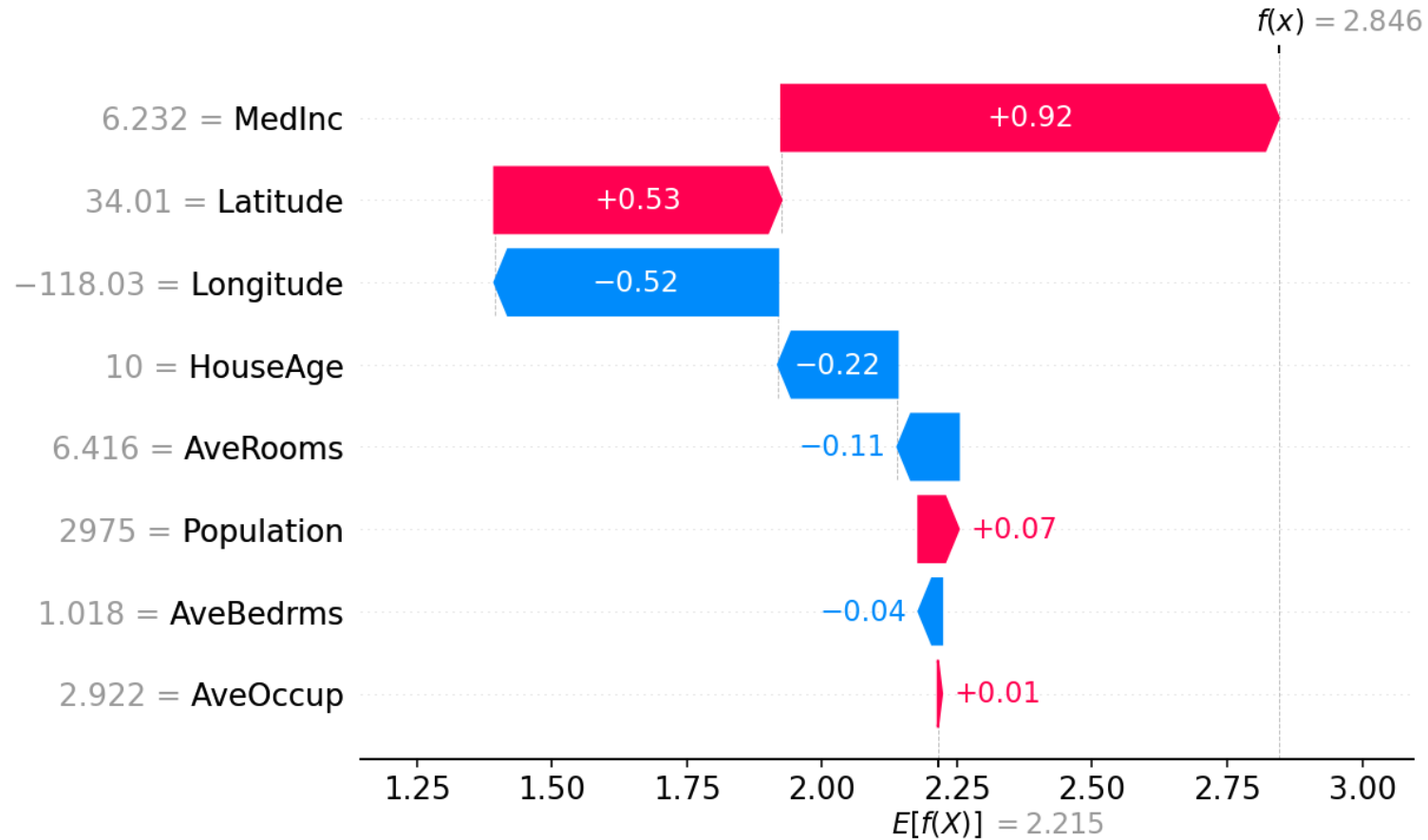


SHAP. График частичной зависимости

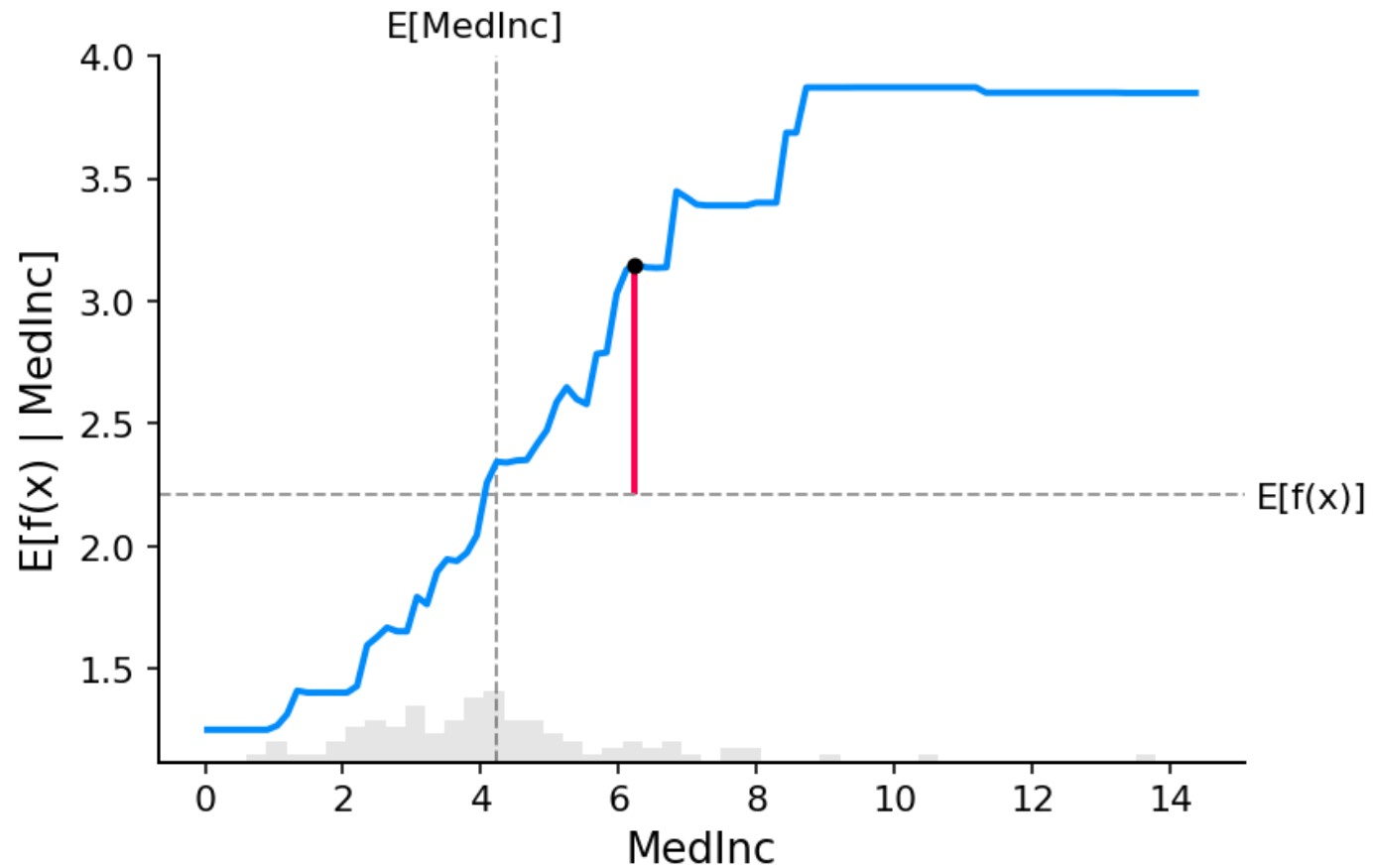


SHAP. Аддитивный характер значений

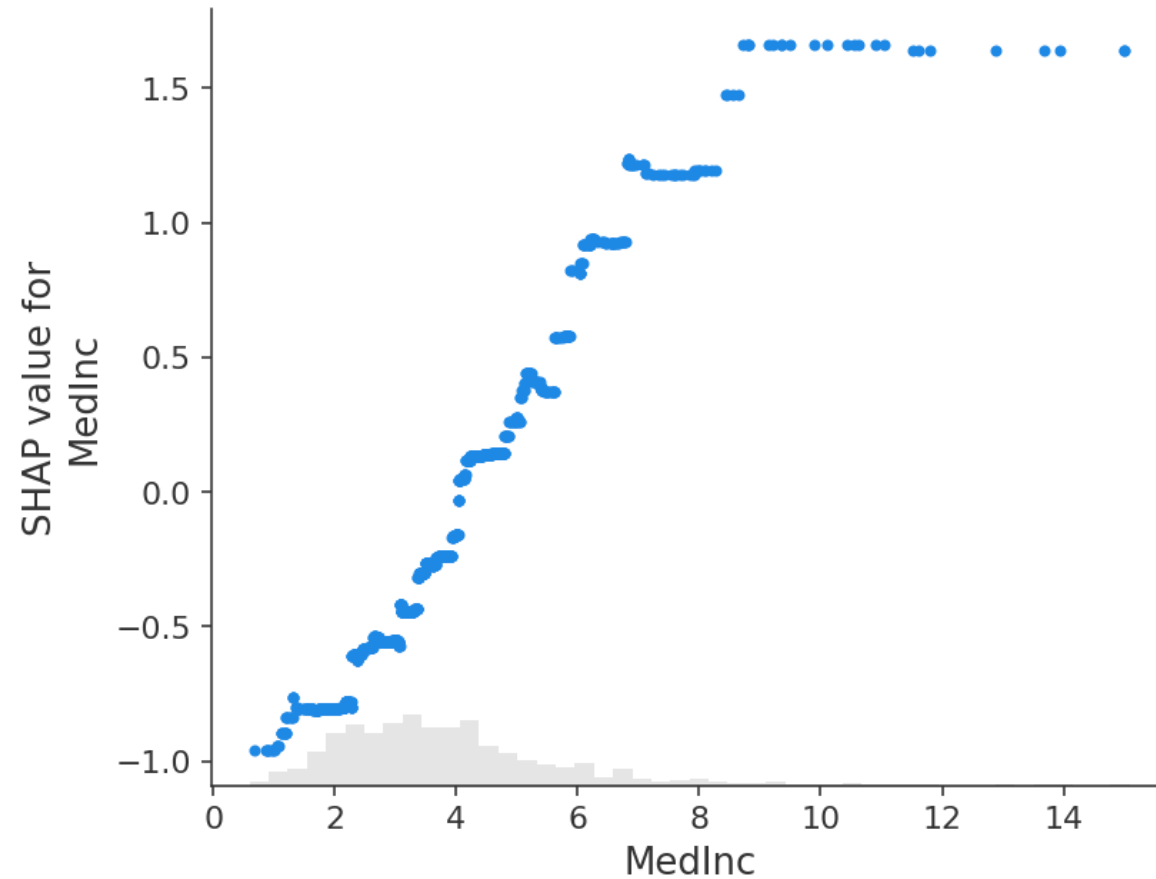
Шепли



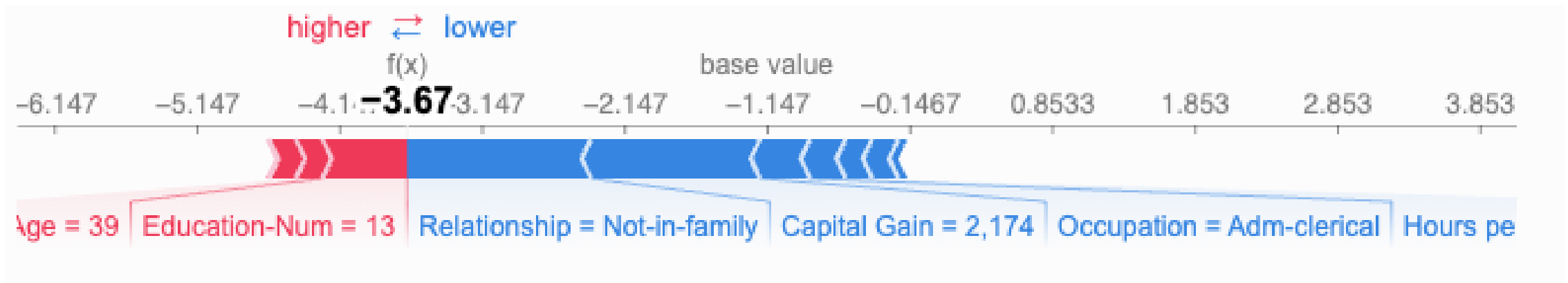
SHAP. График частичной зависимости



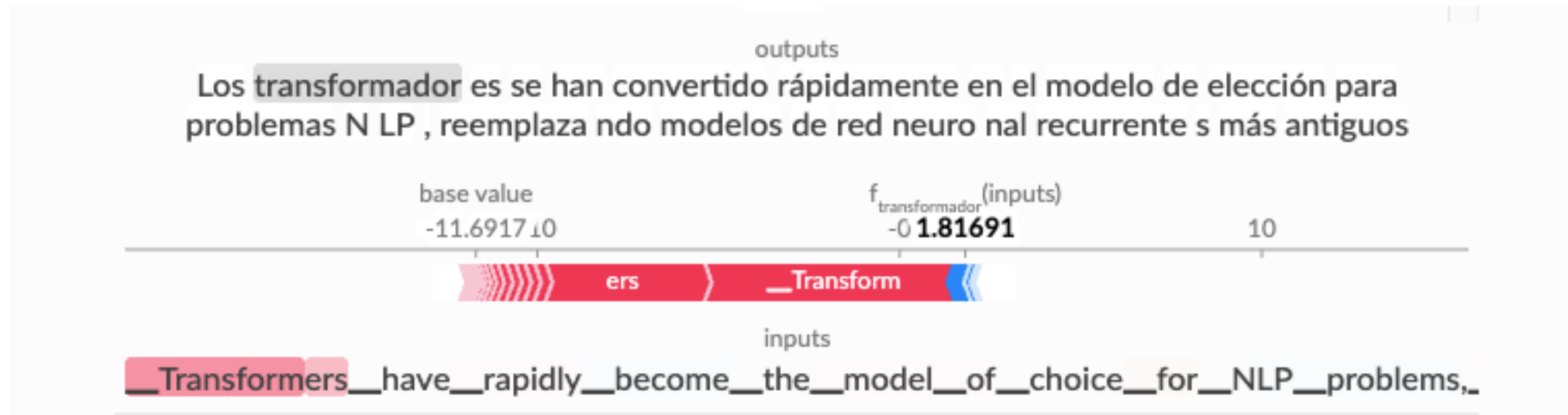
SHAP. График частичной зависимости



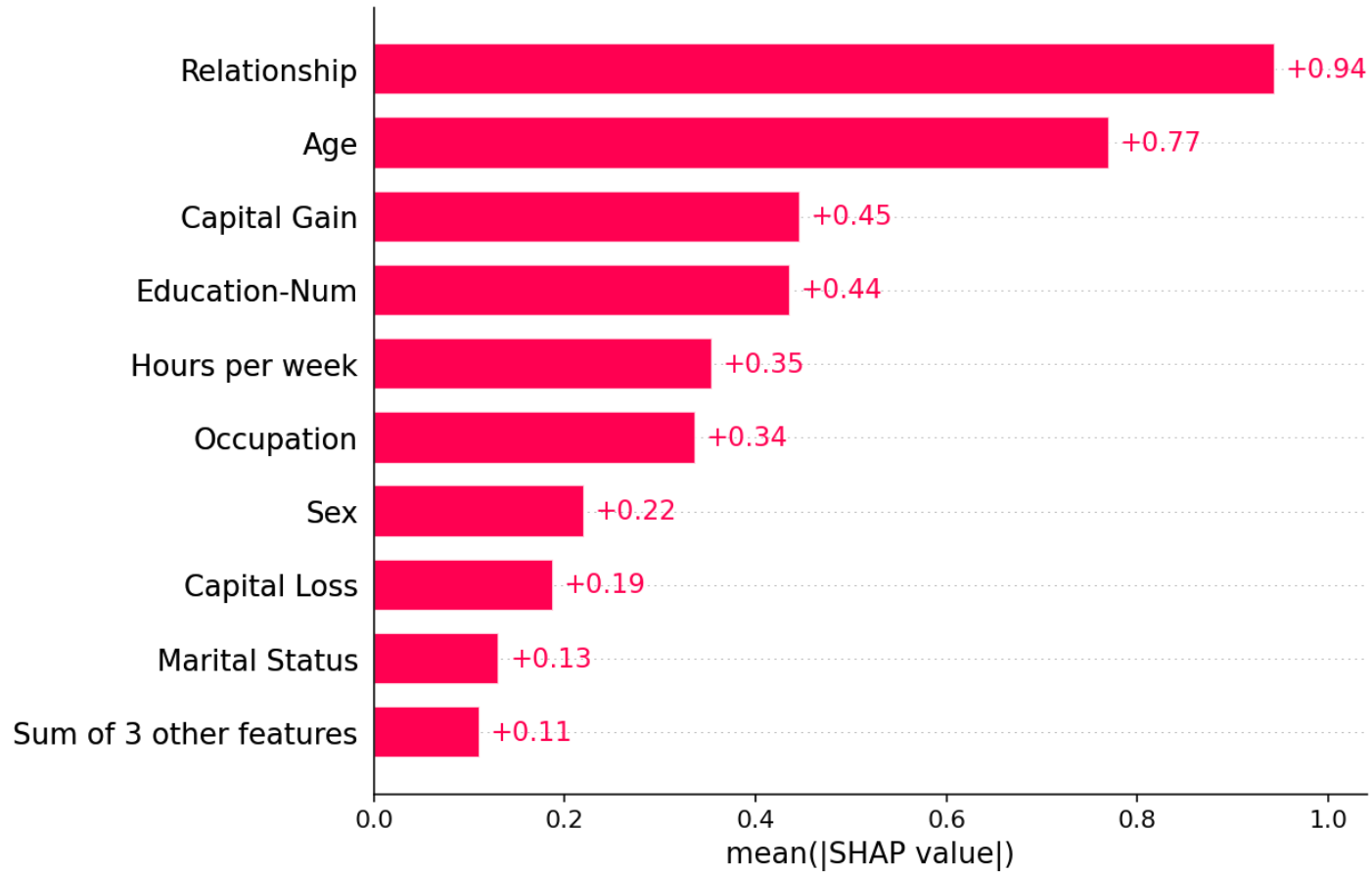
SHAP. Force Plot



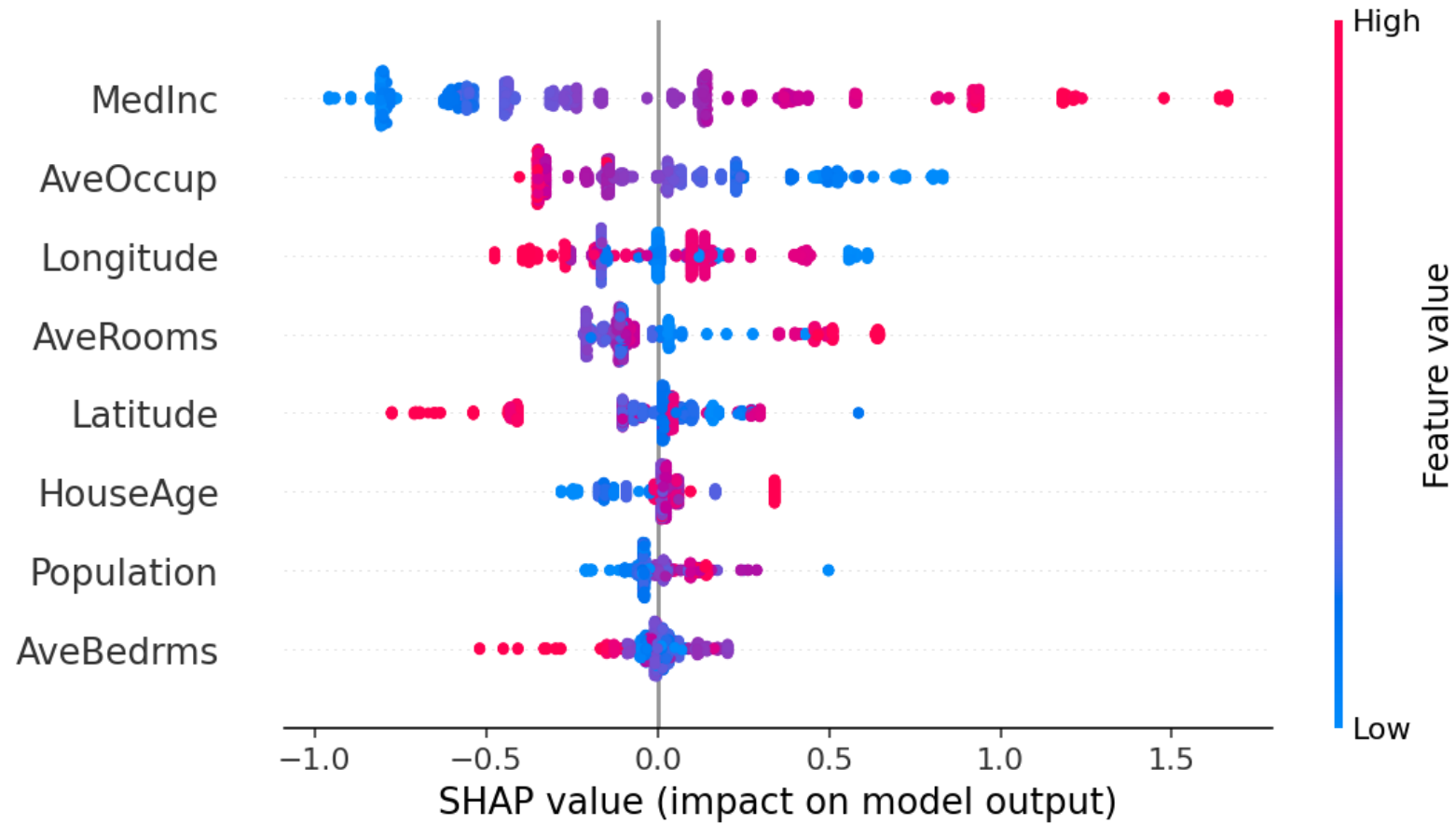
SHAP. Force Plot



SHAP. Bar Plot



SHAP. Beeswarm



SHAP. CV

