Калибровка модели машинного обучения

https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html

https://scikit-learn.org/stable/modules/calibration.html#calibration

метод **predict_proba**
(схожая ситуация с методом **decision_function**)

Может показаться, что для объекта этот метод оценивает вероятность принадлежать классу.
Но это не так.

Predicting Good Probabilities with Supervised Learning, A. Niculescu-Mizil & R. Caruana, ICML 2005

Градиентный бустинг отодвигает значения вероятностей от нуля и единицы
Наивный Байес сдвигает значения вероятностей в сторону нуля или единицы

Нейронные сети и случайный лес умеренно искажают вероятности.

Наши модели выдают ранжировку объектов распознавания, близкую к вероятностям.

Схожесть и различие между
- ранжировкой и
- вероятностью принадлежать классу

Пример.
Вероятность кликнуть на рекламный банер и установить приложение равна 0.8.

Вопрос
Как из ранжировки получить вероятность принадлежать классу

Ответ.
С помощью калибровки модели.

Зачем могут использоваться вероятности принадлежать классу?

Пусть мы оценили чек покупателя и вероятность покупки

Калибровка не предназначена для улучшения качества распознавания.
Она придает алгоритму дополнительные свойства, полезные в некоторых задачах.

isotonic regression

- разработчики строят ее как кусочно-постоянную функцию,
- отсутствует второй шаг, подгонка функцией, обычно подгонка сплайном

Переобучение при калибровке модели
Калибровку рекомендуют проводить на подвыборке, отличной от обучающего множества.

Изотоническую регрессию рекомендуют проводить, если объем множества, предназначенного для калибровки, превышает 1000 наблюдений.
IMHO, это чересчур...

Кросс-валидация при калибровке