Коэффициент Джини (Gini) измеряет степень неоднородности распределения или точность предсказаний модели. В контексте оценки моделей или переменных в анализе данных он применяется для проверки их способности разделять группы, например, хороших и плохих клиентов в кредитном скоринге. Для непрерывных данных вычисление коэффициента Джини обычно связано с построением ROC-кривой (Receiver Operating Characteristic), поскольку он тесно связан с площадью под этой кривой (AUC — Area Under the Curve).

1. **Вычисление на основе ROC-кривой**. Для оценки Джини переменной или модели сначала необходимо получить предсказанные значения (например, вероятность принадлежности к классу) и реальное распределение классов. Затем строится ROC-кривая, которая отображает соотношение между долей истинных положительных результатов (TPR) и долей ложных положительных результатов (FPR) при различных порогах классификации. Коэффициент Джини связан с AUC формулой:

* Здесь AUC — это площадь под ROC-кривой, которая измеряет, насколько хорошо модель ранжирует объекты.

1. **Оценка для переменных**. Для отдельной переменной необходимо получить ее способность различать классы, например, с помощью подсчёта частот попадания значений переменной в соответствующие классы. Это можно сделать, построив таблицу распределения или бинирование значений переменной и анализируя, насколько они ассоциируются с классами. ROC-кривая строится по этим данным, и далее рассчитывается коэффициент Джини, как описано выше.
2. **Практическое использование и интерпретация**. Коэффициент Джини варьируется от -1 (перевернутое ранжирование) до 1 (идеальное ранжирование), где 0 означает отсутствие дискриминации. В реальных задачах значения выше 0.3–0.4 уже считаются приемлемыми для моделей. Чтобы интерпретировать Джини для модели или переменной, важно учитывать контекст и область применения, так как требования к точности могут различаться.

Для расчета **AUC** (площади под ROC-кривой) можно использовать несколько подходов. Основной метод — это вычисление площади с использованием метода трапеций, что дает приближение для интеграла под кривой. Формула выглядит следующим образом:

Где:

* — значение False Positive Rate (доля ложных срабатываний) на -м шаге,
* — значение True Positive Rate (доля истинных срабатываний) на -м шаге,
* — количество точек, используемых для построения ROC-кривой.

### Расчет AUC через сортировку и ранжирование

Если у вас есть предсказанные вероятности и метки классов, AUC также можно вычислить через сравнение пар значений. Формула для этого подхода:

Для каждой пары , где принадлежит положительному классу, а — отрицательному, проверяется, насколько предсказанные вероятности соответствуют истинному порядку. Если вероятность для положительного больше, пара учитывается.

### Учет связи с коэффициентом Джини

После получения AUC можно вычислить коэффициент Джини для модели или переменной по формуле:

Эти формулы позволяют оценить, насколько хорошо модель или переменная разделяют классы. Чем выше , тем лучше работает модель, а коэффициент Джини отражает эту способность в масштабе от -1 до 1.