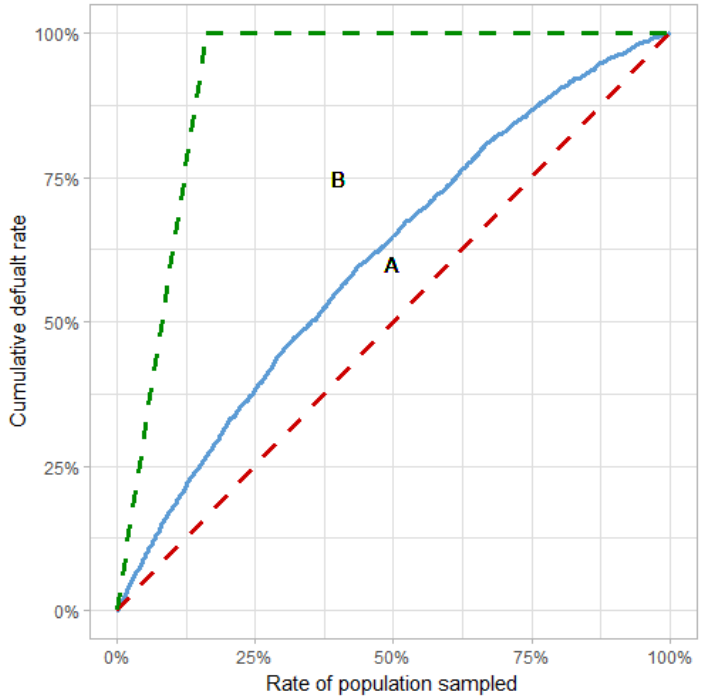
Gini coefficient – Regression (1)

- The Gini coefficient is a metric that indicates the model's discriminatory power, namely, the effectiveness of the model in differentiating between "class 1", and "class 2".
- This metric is often used to compare the quality of different models and evaluate their prediction power.

Example: Distinguish good borrowers from those who default.

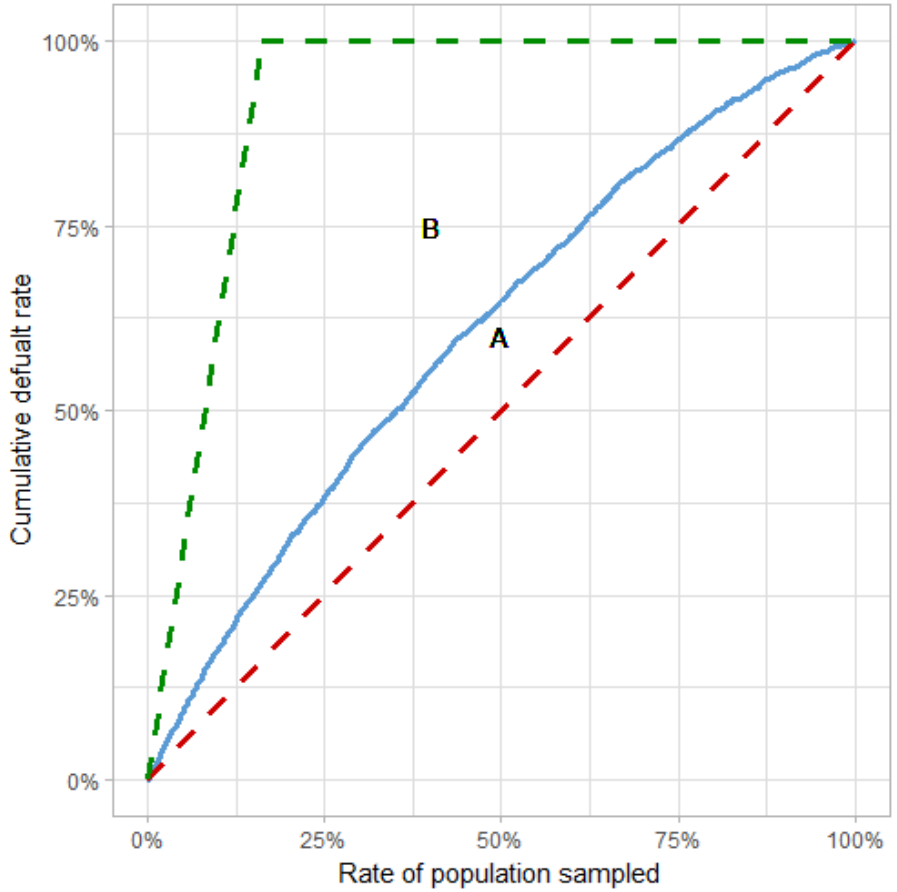
Construct a CAP curve (captures the ordinal relation between score (PD) and default rate): all the model's population needs to be ordered by the predicted likelihood of default. Namely, the observation with the lowest score is first, and the observation with the highest score is last. Then, we sample the population from first to last, and after each sampling, we calculate the cumulative default rate. The x-axis of the CAP curve represents the portion of the population sampled, and the y-axis represents the corresponding cumulative default rate.



https://towardsdatascience.com/using-the-gini-coefficient-to-evaluate-the-performance-of-credit-score-models-59fe13ef420

Gini coefficient – Regression (2)

- Gini index based on the CAP curve:
- If our model has perfect discriminatory power, we would expect to reach 100% of the cumulative default rate after sampling a portion of observations, which is equal to the default rate in our data (the green line in the chart below). E.g., if the default rate in our data is 16%, after sampling 16% of the observations, we would capture all the defaults in our data. On the contrary, if we use a random model, i.e., a model which randomly assigns scores in equal distribution, the cumulative default rate will always equal to the portion of observations sampled (the red line in the chart below).
- The Gini coefficient is defined as the ratio between the area within the model curve and the random model line (A) and the area between the perfect model curve and the random model line (A+B).
- Put it differently, the Gini coefficient is a ratio that represents how close our model to be a "perfect model" and how far it is from being a "random model." Thus, a "perfect model" would get a Gini coefficient of 1, and a "random model" would get a Gini coefficient of 0.



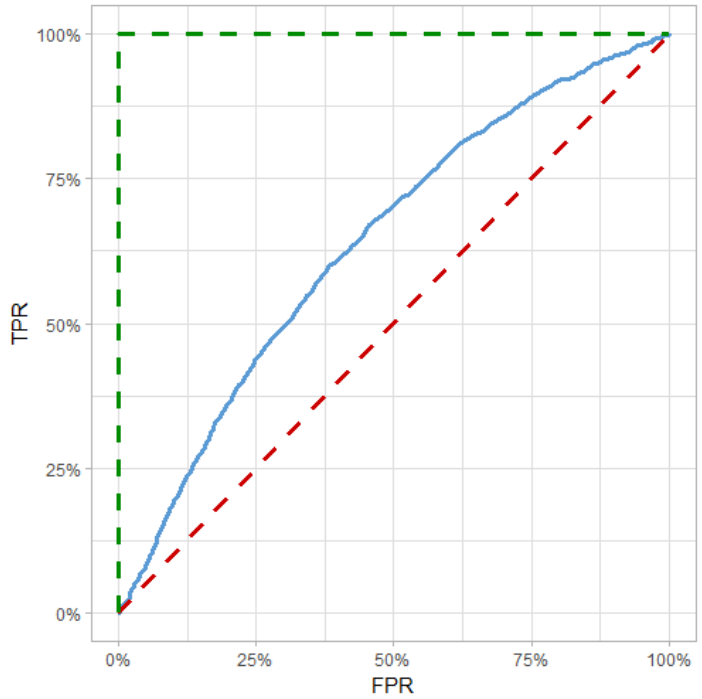
https://towardsdatascience.com/using-the-gini-coefficient-to-evaluate-the-performance-of-credit-score-models-59fe13ef420

Gini coefficient – Regression (3)

- Gini Coefficient can be extracted from the RoC curve

The ROC curve is constructed using confusion matrices that originated from thresholds between 1 to 1000 (in the example below) and plotting their TPR against FPR. The y-axis of the ROC curve represents the TPR values, and the x-axis represents the FPR values. The AUC is the area between the curve and the x-axis.

$$Gini = 2 * AUC - 1$$



https://towardsdatascience.com/using-the-gini-coefficient-to-evaluate-the-performance-of-credit-score-models-59fe13ef420

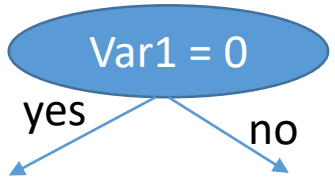
A comprehensive version of computing Gini index is available here: https://link.springer.com/article/10.1057/dbm.2010.2

Gini index – Decision Trees

Class	Var1	Var2
A	0	33
A	0	54
A	0	56
A	0	42
A	1	50
B	1	55
B	1	77
B	0	49

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Gini: 1- (25/64) – (9/64)
= 15/32



A: 4
B: 1

A: 1
B: 2

Gini: 1- (16/25) – (1/25)
= 8/25

Gini: 1- (1/9) – (4/9)
= 4/9

Gini at level 1 = (5/8)*(8/25) + (4/8)*(4/9)
= 1/5 + 2/9
=19/45

Information gain = (15/32) – (19/45) = 0.04653

http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/

Gini index – Contingency table

Gini index (G)	$\max\Big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2] - P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2] - P(A)^2 - P(\overline{A})^2\Big)$

http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/

Which of these is important for us?

- Being aware Gini coefficient can be computed from RoC (2AUC – 1)

and

- Being able to compute the Gini index for Decision Trees
  + Information Gain