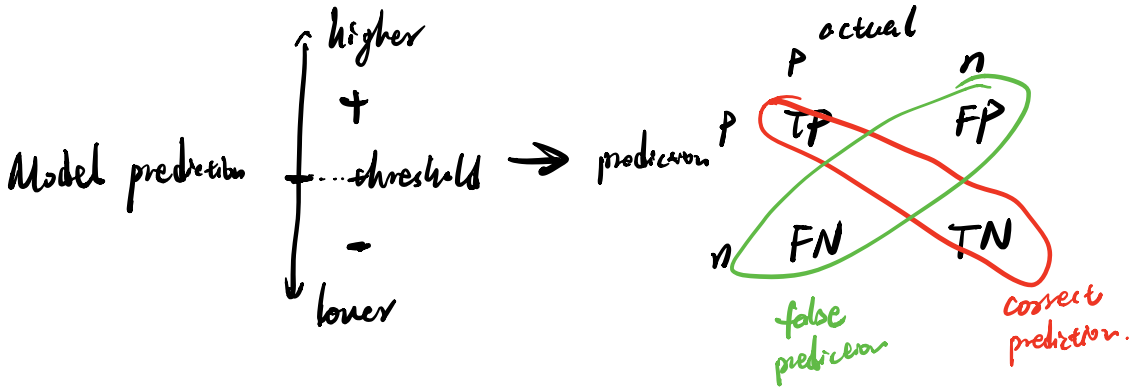
we have the following design options:

[Algorithm. Feature Set. Hyper-parameters (complexity)]

we also need to choose an evaluation metric



Most commonly used:

Accuracy: $(TP+TN)/(TP+FP+TN+FN)$

Precision: $(TP)/(TP+FP)$

Recall: $(TP)/(TP+FN)$

They are dependent on the threshold used for the classifier. you can improve one (but not all) by moving threshold.

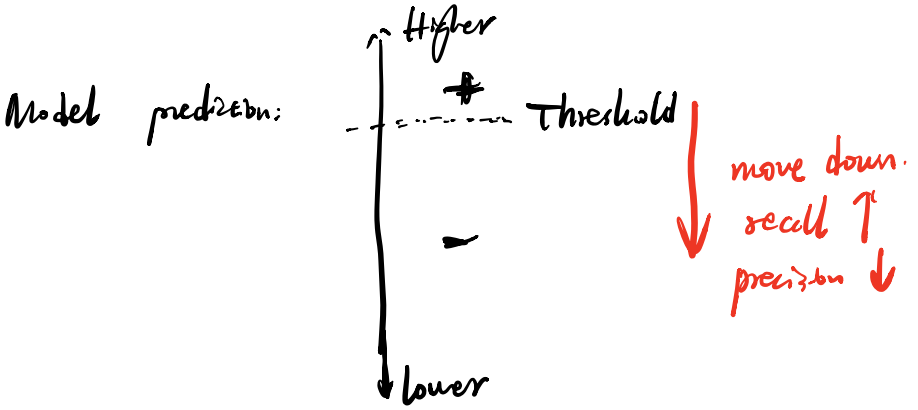
They are also dependent on the base rate p.

if p is 0.05. if you guess everything is negative then accuracy will be 95%., if you guess everything is positive. then recall will be 100%.

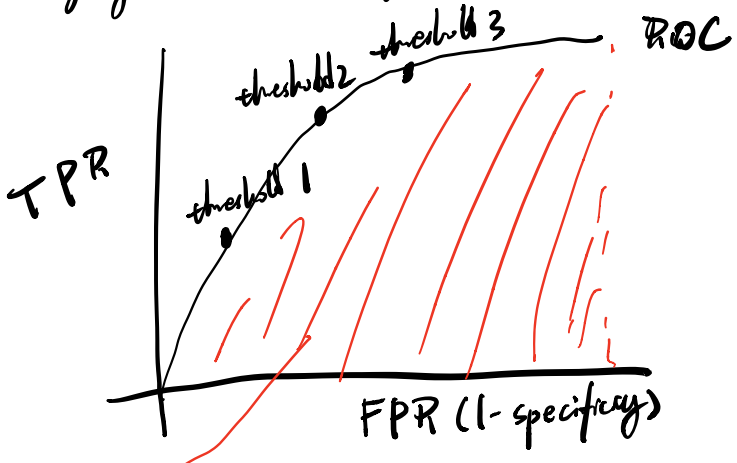
we need to compare them with the baseline.

If there's a good reason to favor both precision and recall. one could use F1-score, the harmonic mean of the two:

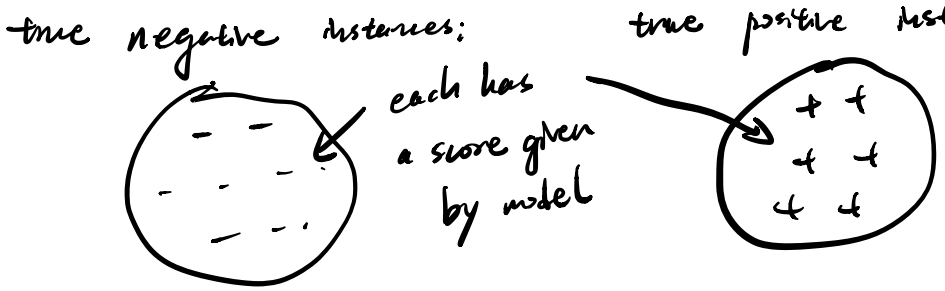
$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



Each threshold we choose creates a trade-off between False positive rate (1- specificity) and True positive rate. so changing threshold gives us a curve.



AUC helps us select the model ($0.5 \leq AUC \leq 1$)

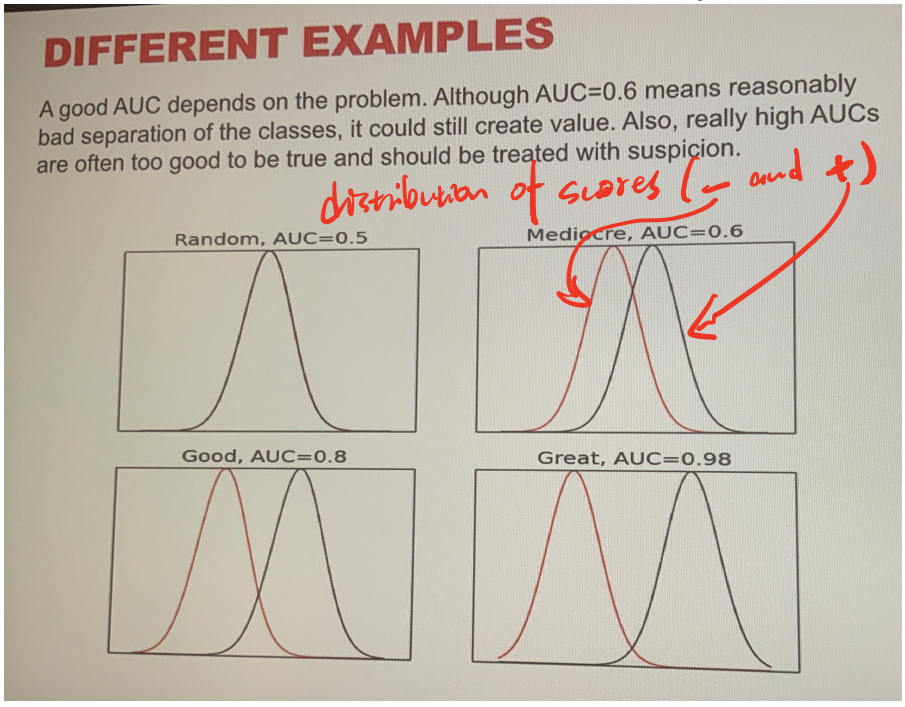


you randomly draw a positive instance. and randomly draw a negative instance. AUC is the probability that the score of the randomly drawed positive is higher than the score of the randomly drawed negative.

$$AUC = P(f(x^+) > f(x^-))$$

$0.5 < AUC < 1$

↑ random guessing ↑ perfect seperation



FUN AUC FACTS

$P(f(x) > f(-))$

- **Nice interpretation**: gives the probability that a positive instance will have a higher score than a negative instance

- **Base Rate Invariant**: AUC is invariant to P(+) in the data set (unlike other classification metrics). Useful for doing comparisons across data sets with different base rates. Or after down sampling.

$[0.5, 1]$

- **Is Nicely Bounded:** AUC scores range from [0,1], where 1 is a perfect classifier and 0 is a perfectly wrong classifier. A random classifier has an exact score of 0.5.

LIFT:

The ratio of target response to the average response

ex: Suppose an average response rate is 5%. but the model identifies a segment with a response rate of 20%. then lift is $\frac{20\%}{5\%} = 4$

Typically. the modeller seeks to divide the population into quantiles. and rank the quantiles by lift.

segment of population

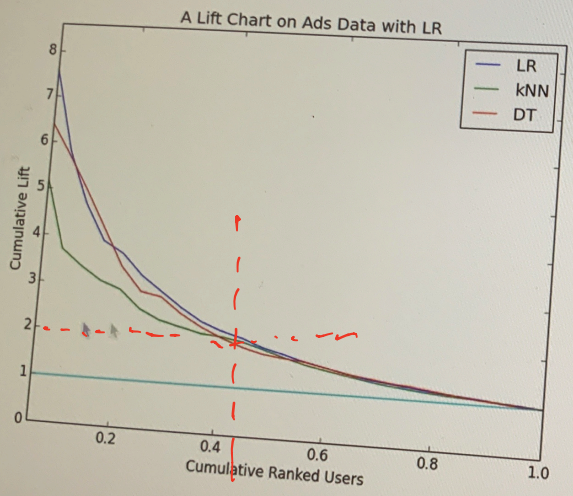
$$Lift(B \Rightarrow 1) = \frac{P(1|B)}{P(1)} = \frac{P(1 \wedge B)}{P(B)\,P(1)}$$

$$conf(B \Rightarrow 1) = P(1|B)$$

$$\text{so } lift(B \Rightarrow 1) = \frac{conf(B \Rightarrow 1)}{P(1)}$$

LIFT

Lift can be both a ranking metric and a classification metric. For ranking, we can see which model fits the entire distribution of users better. For single classification, we can measure lift for a desired targeting threshold.



Lift Properties

- **Nice interpretation**: the lift tells you exactly how many more positive outcomes you might expect relative to the baseline strategy. Also lends well to economic analysis

- **Not Base Rate Invariant:** Lift will change if you alter P(+). This has implications for down sampling or for comparing models from different datasets.

Lift depends on the base rate.

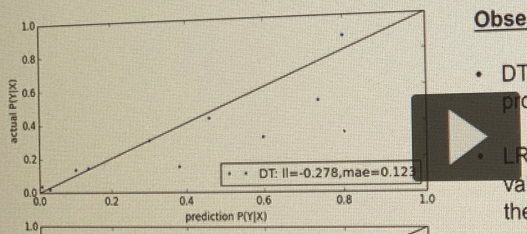
if base rate = 50%. the most lift is 2 (100%/50%=2)

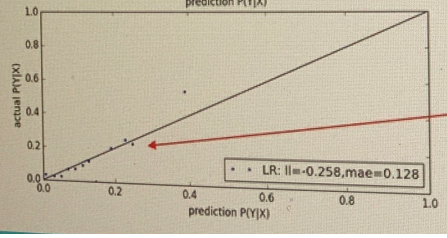
if base rate = 5%, the most lift can be 20

CALIBRATION PLOTS

we want our model's predicted probability is the same as the actual probability

A good way to test our ability to estimate probabilities is to generate calibration plots. To make the plot, we bin test instances by P(Y|X) and take mean(Y) against mean(P(Y|X)) for each bin.



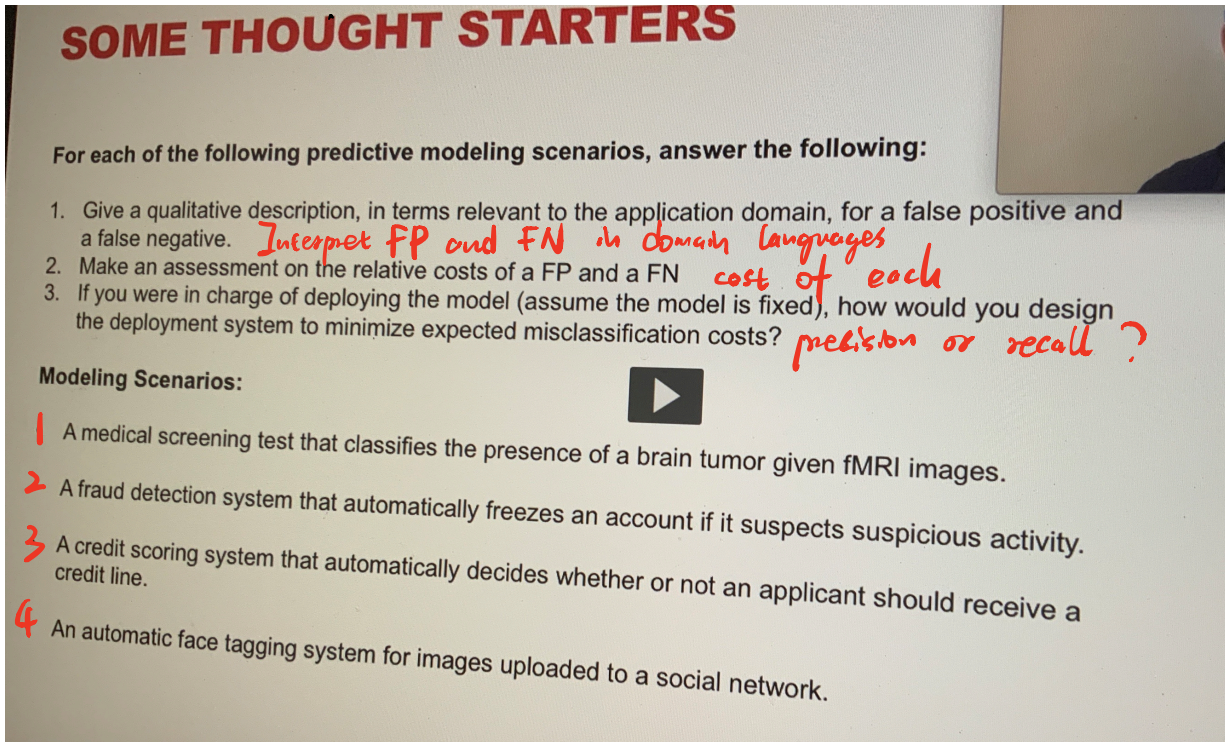


Observations for this problem:

- DT predicts a higher range of probabilities
- LR is very well calibrated for lower valued predictions but not as good in the upper range.

We want to see our points line up against the identity line as much as possible.

Decision Systems:



Interpret FP and FN in domain languages

cost of each

precision or recall ?

1. FP: we think the person has a tumor, but he doesn't

FN, we think the person does not has a tumor, but he does.

cost of FN >> cost of FP

cost of FN: Test N, then the person very likely won't check again, and the tumor grows, leading to a person's death.

Model design: lower the threshold to increase the Recall (decrease the FN rate).

2. FP: we think the activity is fraud. but it's not

FN: we think the activity is not fraud but it is.

cost of FN >> cost of FP

cost of FN: the person could lost money, losing the customers.
if credit card company. lose money as well.

cost of FP, making a confirmation phone call. and no more.

Model design: lower the threshold. increase the Recall.

3. FP: we give credit to user, but he should not be given.

FN: we don't credit to user when he's actually good.

cost of FP >> cost of FN

cost of FP: hurt the company's money. hurt the user's credit score, and even break down economy.

Model design: Increase the threshold. making the model more conservative, and increase the precision