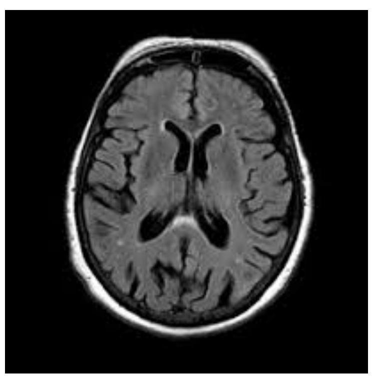
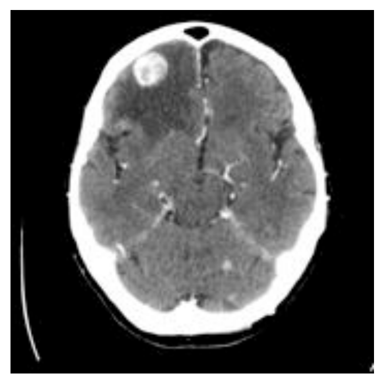
DAT 13, Imbalanced Classes / ROC Curves

Your are a data scientist for a project called CancerScreen. You are tasked with creating a new classifier that classifies radiology images of brains as having cancer or not having cancer.

In this project, all the images that have been tagged as cancerous will go to a trained physician for further review



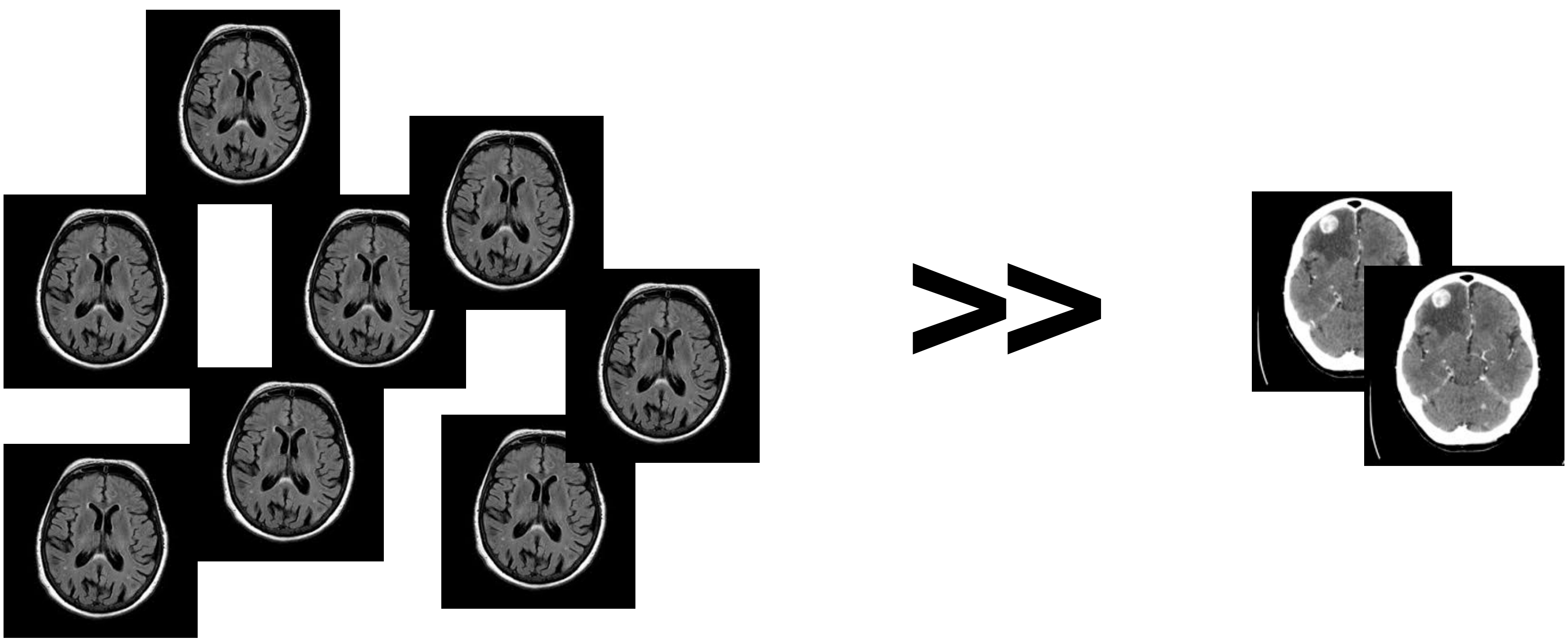
No Cancer



Cancer

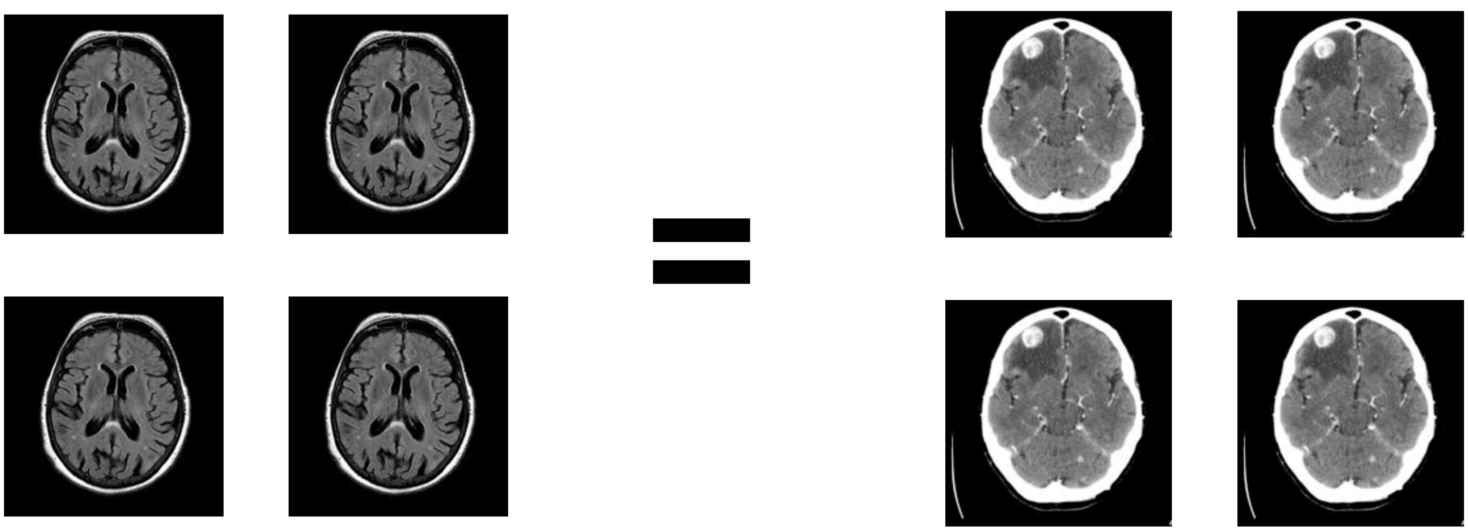
First Issue: There are a lot more healthier brains than cancerous brains. This imbalance will confuse many classifiers as they will only perform well on the dominant class and poorly on the minority class.

This situation shows up frequently in many fields ex. fraud detection, medical diagnosis, etc.



Solution: Balance your classes and train on this balanced dataset. This can be done by:

1. Undersampling the dominant class - remove some the majority class so it has less weight
2. Oversampling the minority class - add more of the minority class so it has more weight.
3. Hybrid - doing both



Details / Drawbacks

Undersampling — Randomly remove elements from the majority class.

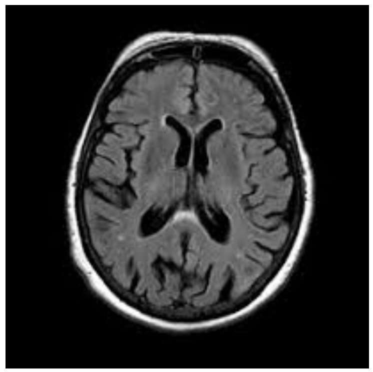
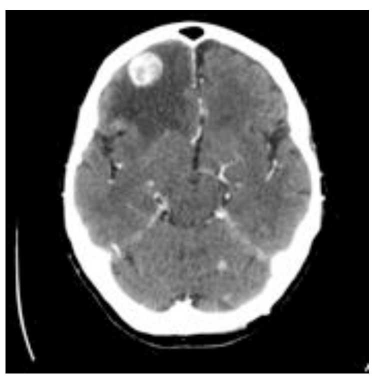
Drawback: Removing datapoints could lose important information

Oversampling — Duplicate elements of your minority class

Drawback: Just replicating randomly minority classes could cause overfit

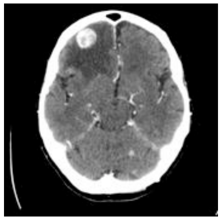
Smote is a more intelligent solution for the oversampling problem

Second Problem: Not all errors are equal …

Error 1		Classifier Label: Cancerous	Permissable, because a physician will review it
Error 2		Classifier Label: Non-Cancerous	Not permissable, because this data will be discarded

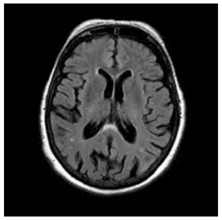
So we need a more sophisticated model of Error Rate:

TP: An Example that is **positive** and is classified as **positive**



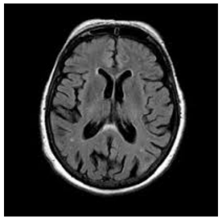
Label: Positive

TN: An Example that is **negative** and is classified as **negative**



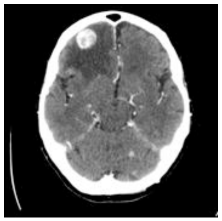
Label: Negative

FP: An Example that is **negative** and is classified as **positive**



Label: Positive

FN: An Example that is **positive** and is classified as **negative**

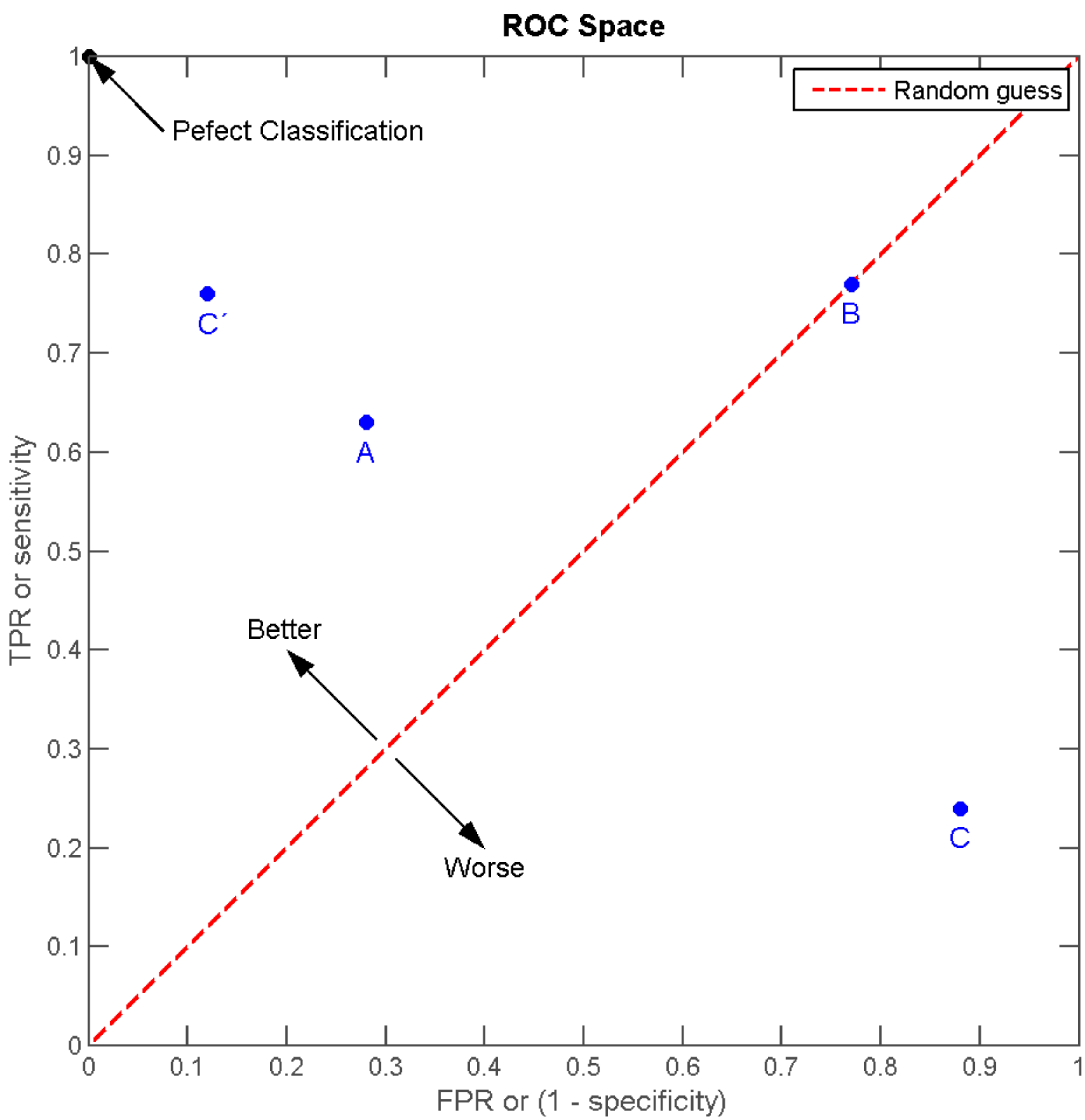


Label: Negative

Metrics

Name	Formula	Explanation
True Positive (TP) Rate	TP / (TP + FP)	The closer to 1, the better. TP rate = 1 when FP =
True Negative (TN) Rate	TN / (TN + FN)	The closer to 1 the better. TN Rate = 1 when FN = 0
False Positive (FP) Rate	FP / (FP + TN)	The closer to 0 the better when FP rate = 0 when FP =0
False Negative Rate / (FN + TP)	FN / (FN + TP)	The closer to 0 the better. FN rate = 0 when FN = 0

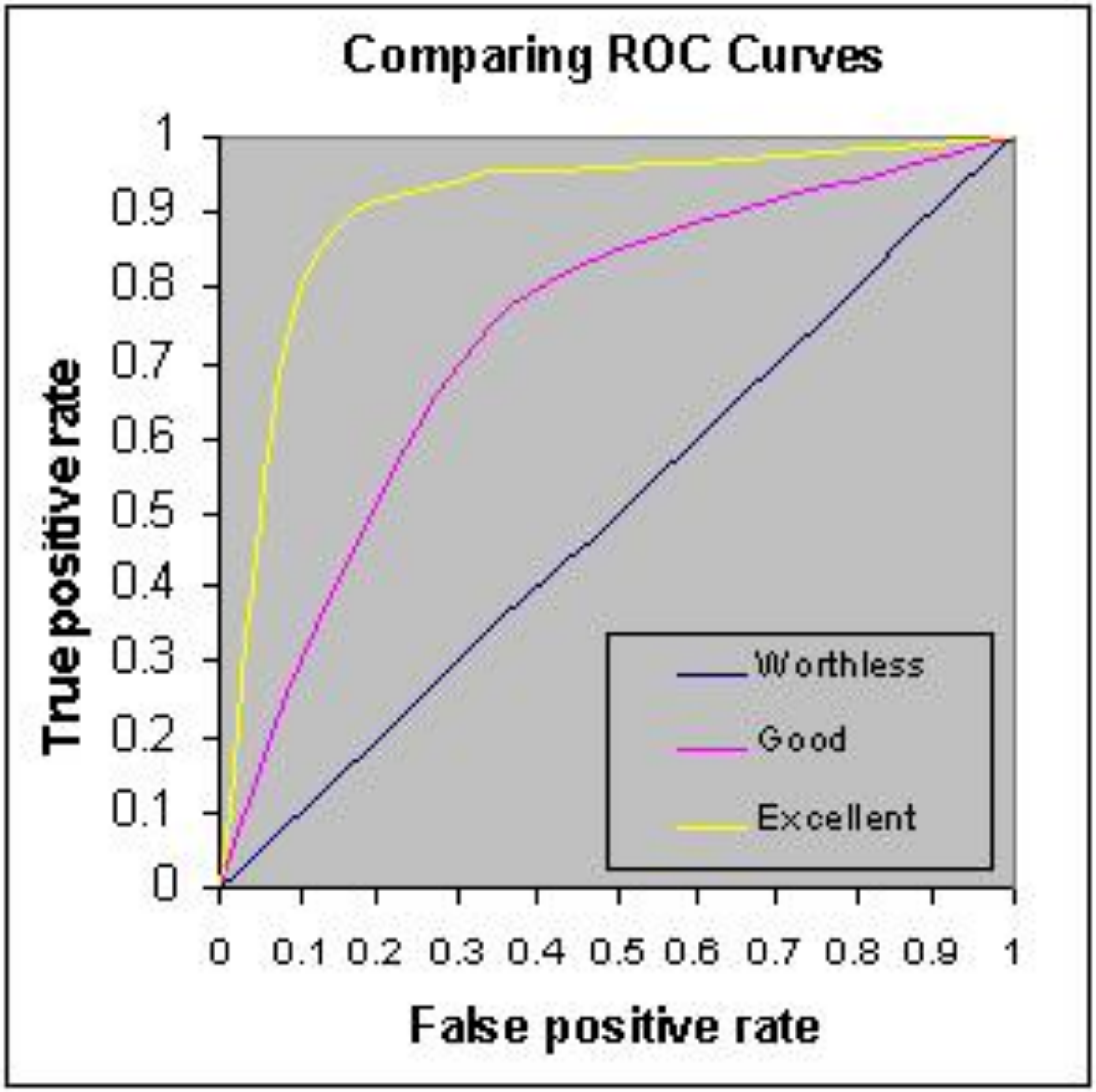
Let’s first understand ROC Space ….



TP Rate = True Positives / All positives

FP Rate = False Positives / All Negatives

ROC Curves show the relationship between the TP Rate and the FP Rate as we vary the decision threshold for the classifier



Evaluating A Classifier using ROC

We evaluate a classifier by measuring the area under the curve for its ROC curve. The Greater area under the curve, the more effective the classifier.

Then for our chosen classifer, we pick an appropriate decision threshold. In general, we pick the decision threshold that gets us closest to the upper left corner

Review for Imbalanced Classes

1. Balance your dataset so that the number of elements in each class are equal
2. Train different classifiers on this balanced data
3. For each classifier, create an ROC curve and compute the Area under the Curve (AUC) for each one
4. For the classifier with the greatest AUC, pick the appropriate decision threshold given the specifics of your problem

Precision/Recall - Another Measure of Performance

Counts: Lions: 4, Tigers: 5, House Cats: 4

Total: 13

We've trained our classifier on tigers and ask it to find all the tigers in this dataset. Here's what it returns:

Classified as Tigers:






Precision is the percentage of True Positives in your set of results

= 4 / 6 = .66

Recall is True Postives / Total Positives.
Same as TPR

= 4 / 5 = .8

Formulas

$$\begin{aligned}
\text{Recall} &= \frac{TP}{TP+FN} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{True Positive Rate} &= \frac{TP}{TP+FN} \\
\text{False Positive Rate} &= \frac{FP}{FP+TN}
\end{aligned}$$