# Which Classifer is better? High Skew case

| T1 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 50 | 50 |
| | Class=No | 100 | 9900 |

Precision $(p) = 0.3$

TPR $=$ Recall $(r) = 0.5$

FPR $= 0.01$

TPR/FPR $= 50$

F $-$ measure $= 0.375$

| T2 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 1000 | 9000 |

Precision $(p) = 0.09$

TPR $=$ Recall $(r) = 0.99$

FPR $= 0.1$

TPR/FPR $= 9.9$

F $-$ measure $= 0.165$

| T3 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 100 | 9900 |

Precision $(p) = 0.5$

TPR $=$ Recall $(r) = 0.99$

FPR $= 0.01$

TPR/FPR $= 99$

F $-$ measure $= 0.66$

## Building Classifiers with Imbalanced Training Set

- Modify the distribution of training data so that rare class is well-represented in training set
  - Undersample the majority class
  - Oversample the rare class

# Which Classifer is better?

| T1 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 50 | 50 |
| | Class=No | 1 | 99 |

Precision $(p) = 0.98$

$\text{TPR} = \text{Recall} \ (r) = 0.5$

$\text{FPR} = 0.01$

$\text{TPR/FPR} = 50$

$\text{F} - \text{measure} = 0.66$

| T2 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 10 | 90 |

Precision $(p) = 0.9$

$\text{TPR} = \text{Recall} \ (r) = 0.99$

$\text{FPR} = 0.1$

$\text{TPR/FPR} = 9.9$

$\text{F} - \text{measure} = 0.94$

| T3 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 1 | 99 |

Precision $(p) = 0.99$

$\text{TPR} = \text{Recall} \ (r) = 0.99$

$\text{FPR} = 0.01$

$\text{TPR/FPR} = 99$

$\text{F} - \text{measure} = 0.99$

# Which Classifer is better? Medium Skew case

| T1 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 50 | 50 |
| | Class=No | 10 | 990 |

Precision $(p) = 0.83$

$\text{TPR} = \text{Recall} \ (r) = 0.5$

$\text{FPR} = 0.01$

$\text{TPR/FPR} = 50$

$\text{F} - \text{measure} = 0.62$

| T2 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 100 | 900 |

Precision $(p) = 0.5$

$\text{TPR} = \text{Recall} \ (r) = 0.99$

$\text{FPR} = 0.1$

$\text{TPR/FPR} = 9.9$

$\text{F} - \text{measure} = 0.66$

| T3 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 10 | 990 |

Precision $(p) = 0.9$

$\text{TPR} = \text{Recall} \ (r) = 0.99$

$\text{FPR} = 0.01$

$\text{TPR/FPR} = 99$

$\text{F} - \text{measure} = 0.94$

# Which of these classifiers is better?

| A | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 10 | 40 |
| | Class=No | 10 | 40 |

$\text{Precision} \quad (p) = 0.5$

$\text{TPR} \;= \text{Recall} \quad (r) = 0.2$

$\text{FPR} \;= 0.2$

$\text{F}-\text{measure} = 0.28$

| B | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 25 | 25 |
| | Class=No | 25 | 25 |

$\text{Precision} \quad (p) = 0.5$

$\text{TPR} \;= \text{Recall} \quad (r) = 0.5$

$\text{FPR} \;= 0.5$

$\text{F}-\text{measure} = 0.5$

| C | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 40 | 10 |

$\text{Precision} \quad (p) = 0.5$

$\text{TPR} \;= \text{Recall} \quad (r) = 0.8$

$\text{FPR} \;= 0.8$

$\text{F}-\text{measure} = 0.61$

## Dealing with Imbalanced Classes - Summary

- Many measures exists, but none of them may be ideal in all situations
  - Random classifiers can have high value for many of these measures
  - TPR/FPR provides important information but may not be sufficient by itself in many practical scenarios
  - Given two classifiers, sometimes you can tell that one of them is strictly better than the other
    - C1 is strictly better than C2 if C1 has strictly better TPR and FPR relative to C2 (or same TPR and better FPR, and vice versa)
  - Even if C1 is strictly better than C2, C1's F-value can be worse than C2's if they are evaluated on data sets with different imbalances
  - Classifier C1 can be better or worse than C2 depending on the scenario at hand (class imbalance, importance of TP vs FP, cost/time tradeoffs)

# Alternative Measures

| A | | PREDICTED CLASS | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 10 | 40 |

Precision (p) $= 0.8$

TPR $=$ Recall (r) $= 0.8$

FPR $= 0.2$

F−measure (F) $= 0.8$

Accuracy $= 0.8$

$$\frac{\text{TPR}}{\text{FPR}} = 4$$

| B | | PREDICTED CLASS | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 1000 | 4000 |

Precision (p) $= 0.038$

TPR $=$ Recall (r) $= 0.8$

FPR $= 0.2$

F−measure (F) $= 0.07$

Accuracy $= 0.8$

$$\frac{\text{TPR}}{\text{FPR}} = 4$$

# Measures of Classification Performance

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Yes | No |
| ACTUAL CLASS | Yes | TP | FN |
| | No | FP | TN |

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$ErrorRate = 1 - accuracy$$

$$Precision = Positive\ Predictive\ Value = \frac{TP}{TP + FP}$$

$$Recall = Sensitivity = TP\ Rate = \frac{TP}{TP + FN}$$

$$Specificity = TN\ Rate = \frac{TN}{TN + FP}$$

$$FP\ Rate = \alpha = \frac{FP}{TN + FP} = 1 - specificity$$

$$FN\ Rate = \beta = \frac{FN}{FN + TP} = 1 - sensitivity$$

$$Power = sensitivity = 1 - \beta$$

$\alpha$ is the probability that we reject the null hypothesis when it is true. This is a Type I error or a false positive (FP).

$\beta$ is the probability that we accept the null hypothesis when it is false. This is a Type II error or a false negative (FN).

# Alternative Measures

# Alternative Measures

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 10 | 0 |
| | Class=No | 10 | 980 |

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F-measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 1 | 9 |
| | Class=No | 0 | 990 |

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F-measure (F)} = \frac{2*0.1*1}{1+0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

# Which of these classifiers is better?

A

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 10 | 40 |

$\text{Precision (p)} = 0.8$

$\text{Recall (r)} = 0.8$

$\text{F-measure (F)} = 0.8$

$\text{Accuracy} = 0.8$

B

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 1000 | 4000 |

$\text{Precision (p)} =\sim 0.04$

$\text{Recall (r)} = 0.8$

$\text{F-measure (F)} =\sim 0.08$

$\text{Accuracy} =\sim 0.8$

# Alternative Measures

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | a | b |
| | Class=No | c | d |

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

# Alternative Measures

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 10 | 0 |
| | Class=No | 10 | 980 |

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

## Which model is better?

A

|  | PREDICTED |  |  |
|---|---|---|---|
|  |  | Class=Yes | Class=No |
| ACTUAL | Class=Yes | 0 | 10 |
|  | Class=No | 0 | 990 |

Accuracy: 99%

B

|  | PREDICTED |  |  |
|---|---|---|---|
|  |  | Class=Yes | Class=No |
| ACTUAL | Class=Yes | 10 | 0 |
|  | Class=No | 500 | 490 |

Accuracy: 50%

## Which model is better?

A

|  | PREDICTED |  |  |
|---|---|---|---|
|  |  | Class=Yes | Class=No |
| ACTUAL | Class=Yes | 5 | 5 |
|  | Class=No | 0 | 990 |

B

|  | PREDICTED |  |  |
|---|---|---|---|
|  |  | Class=Yes | Class=No |
| ACTUAL | Class=Yes | 10 | 0 |
|  | Class=No | 500 | 490 |

## Accuracy

| | | PREDICTED CLASS | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | a<br>(TP) | b<br>(FN) |
| | Class=No | c<br>(FP) | d<br>(TN) |

□ Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

## Problem with Accuracy

□ Consider a 2-class problem

- Number of Class NO examples = 990
- Number of Class YES examples = 10

□ If a model predicts everything to be class NO, accuracy is 990/1000 = 99 %

- This is misleading because this trivial model does not detect any class YES example
- Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

| | | PREDICTED CLASS | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 0 | 10 |
| | Class=No | 0 | 990 |

## Class Imbalance Problem

- Lots of classification problems where the classes are skewed (more records from one class than another)
  - Credit card fraud
  - Intrusion detection
  - Defective products in manufacturing assembly line
  - COVID-19 test results on a random sample

- **Key Challenge**:
  - Evaluation measures such as accuracy are not well-suited for imbalanced class

## Confusion Matrix

- Confusion Matrix:

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | a | b |
| | Class=No | c | d |

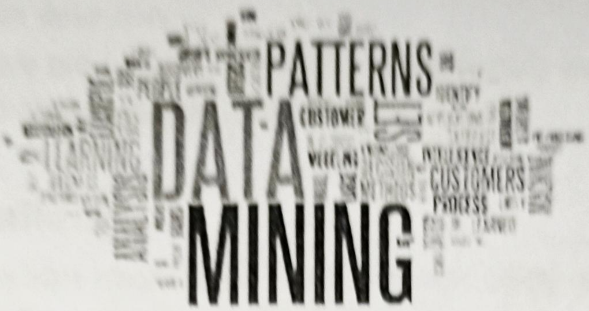
**a: TP (true positive)**

**b: FN (false negative)**

**c: FP (false positive)**

**d: TN (true negative)**

# ITIS404
# Data Mining/Business Intelligence



Spring 2024

## Data Mining
## Classification: Alternative Techniques

Imbalanced Class Problem

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar