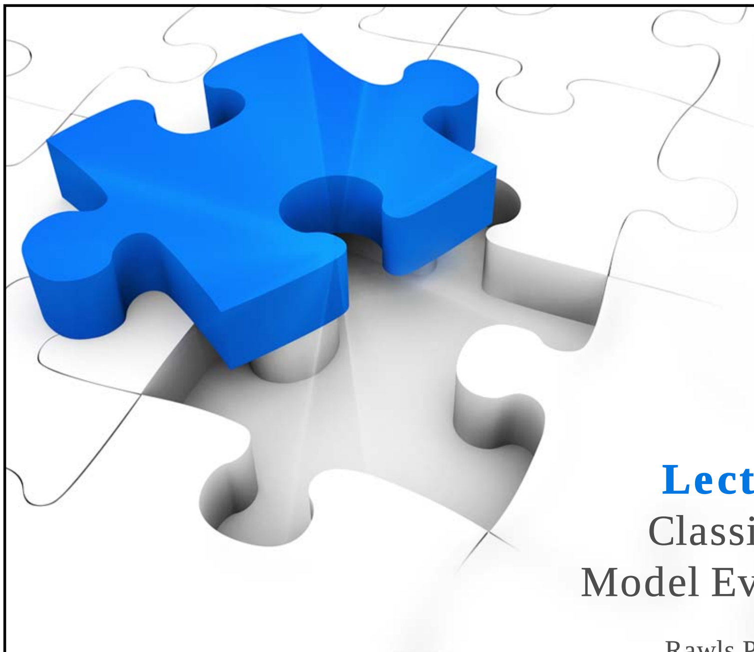


# Lecture 06:
## Classification: Model Evaluation

Rawls Profess of MIS
Jaeki Song, Ph.D.

## Model Evaluation

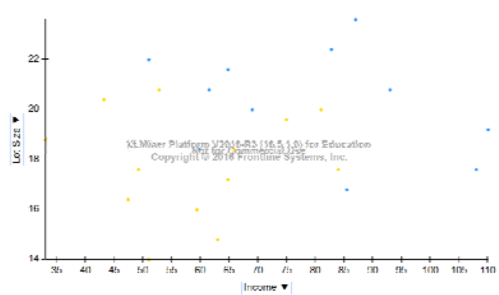
- Supervised Learning
  - interested in predicting the outcome variables for new records
    - predicted numeric value
    - predicted class membership
    - propensity

## Judging Classifier Performance

- The need for performance measures
  - A natural criterion
    - a classifier is the probability of making a *misclassification error*

## Judging Classifier Performance

- Is there a minimal probability of misclassification that we should require of a classifier?
  - benchmark
    - the naïve rule
      - classify as belonging to the most prevalent class
  - class separation



## Model Evaluation Techniques for the Classification Task

- Classification matrix
  - confusion matrix

| | | Predicted Category | | |
|---|---|---|---|---|
| | | 0 | 1 | Total |
| Actual Category | 0 | True Negatives: Predicted 0 Actually 0 | False Positives: Predicted 1 Actually 0 | Total Actually Negative |
| | 1 | False Negatives: Predicted 0 Actually 1 | True Positives: Predicted 1 Actually 1 | Total Actually Positive |
| | Total | Total Predicted Negative | Total Predicted Positive | Grand Total |

| | | Predicted Category | | |
|---|---|---|---|---|
| | | $\le 50K$ | $> 50K$ | Total |
| Actual Category | $\le 50K$ | 18,197 | 819 | 19,016 |
| | $> 50K$ | 2561 | 3423 | 5984 |
| | Total | 20,758 | 4242 | 25,000 |

## Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

$$\text{Error Rate} = \frac{b+c}{a+b+c+d}$$

## Accuracy and Overall Error Rate

- Accuracy
  - represents an overall measure of the proportion of correct classifications being made by the model, while overall error rate measures the proportion of incorrect classifications, across all cells in the contingency table

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = 1 - Error$$

## Computation

| Model $M_1$ | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | + | - |
| | + | 150 | 40 |
| | - | 60 | 250 |

Accuracy = 80%

| Model $M_2$ | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | + | - |
| | + | 250 | 45 |
| | - | 5 | 200 |

Accuracy = 90%

## Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

## Limitation of Accuracy

- Propensities and cutoff for classification
  - cutoff value set by the analyst
    - two-class classifier -> default cutoff → 0.5
    - a cutoff can be either higher or lower

| ID | Actual Class | Probability of Class "Owner" | ID | Actual Class | Probability of Class "Owner" |
|---|---|---|---|---|---|
| 1 | Owner | 0.9959 | 13 | Owner | 0.5055 |
| 2 | Owner | 0.9857 | 14 | Nonowner | 0.4713 |
| 3 | Owner | 0.9844 | 15 | Nonowner | 0.3371 |
| 4 | Owner | 0.9804 | 16 | Owner | 0.2179 |
| 5 | Owner | 0.9481 | 17 | Nonowner | 0.1992 |
| 6 | Owner | 0.8892 | 18 | Nonowner | 0.01494 |
| 7 | Owner | 0.8476 | 19 | Nonowner | 0.0479 |
| 8 | Nonowner | 0.7628 | 20 | Nonowner | 0.0383 |
| 9 | Owner | 0.7069 | 21 | Nonowner | 0.0246 |
| 10 | Owner | 0.6807 | 22 | Nonowner | 0.0218 |
| 11 | Owner | 0.6563 | 23 | Nonowner | 0.0161 |
| 12 | Nonowner | 0.6224 | 24 | Nonowner | 0.0031 |

## Limitation of Accuracy

- Why would we want to use cutoff value different from 0.5 if they increase the misclassification rate?

## Sensitivity and Specificity

- Sensitivity measures the ability to detect the important class members correctly
  - the percentage of C1 members classified correctly

$$\text{Sensitivity} = \frac{a}{a+b}$$

- Specificity measures the ability to rule out C2 members correctly
  - the percentage of C2 members classified correctly

$$\text{Specificity} = \frac{d}{c+d}$$

## False Positive Rate and False Negative Rate

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

- *False Positive Rate* and *False Negative Rate* are additive inverses of sensitivity and specificity

$$\text{False Positive Rate} = 1 - \text{Specificity} = \frac{c}{c+d} = \frac{FP}{FP+TN}$$

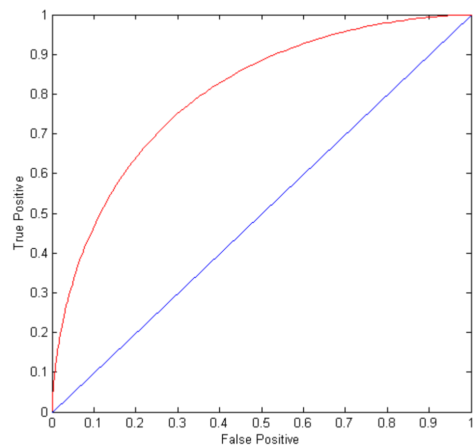
$$\text{Flase Negative Rate} = 1 - \text{Sensitivity} = \frac{b}{a+b} = \frac{FN}{TP+FN}$$

## ROC (Receiver Operating Characteristic)

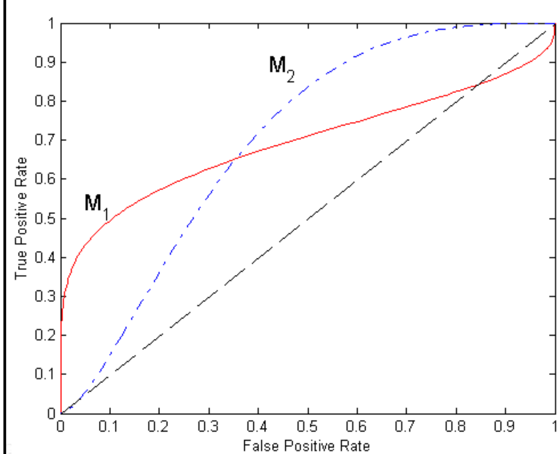
- Developed in 1950s for signal detection theory to analyze noisy signals
  - Characterize the trade-off between positive hits and false alarms
- ROC curve plots sensitivity (on the y-axis) against False Positive Rate (1-specificity) (on the x-axis)
- Performance of each classifier represented as a point on the ROC curve
  - changing the threshold of algorithm, sample distribution or cost matrix changes the location of the point

## ROC Curve

- A common metrics
  - area under curve (AUC) which ranges from 1 (perfect discrimination between classes) to 0.5 (no better that the naïve rule)




## Using ROC for Model Comparison




- No model consistently outperform the other
  - $M_1$ is better for small FPR
  - $M_2$ is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

# How to construct an ROC curve

| Threshold | a | b | c | d | sensitivity | specificity | 1-specificity |
|---|---|---|---|---|---|---|---|
| P ≥ 0.006 | 4 | 0 | 5 | 0 | 1 | 0 | 1 |
| P ≥ 0.026 | 4 | 0 | 4 | 1 | 1 | 0.2 | 0.8 |
| P ≥ 0.1 | 4 | 0 | 3 | 2 | 1 | 0.4 | 0.6 |
| P ≥ 0.103 | 4 | 0 | 2 | 3 | 1 | 0.6 | 0.4 |
| P ≥ 0.356 | 3 | 0 | 1 | 4 | 1 | 0.8 | 0.2 |
| P ≥ 0.728 | 3 | 1 | 1 | 4 | 0.75 | 0.8 | 0.2 |
| P ≥ 0.921 | 2 | 1 | 0 | 5 | 0.75 | 1 | 0 |
| P ≥ 0.981 | 1 | 2 | 0 | 5 | 0.5 | 1 | 0 |
| P ≥ 1.000 | 0 | 3 | 0 | 5 | 0.25 | 1 | 0 |