Classification:
Performance Metrics & Class Imbalance

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Recap
2. Confusion Matrix
   - Accuracy
   - TNR
   - TNR
3. ROC curve
4. Imbalanced Classification
   - Metrics
   - Class rebalancing

Agenda

Classification: Motivation

- Many predictive questions are about classification
  - Credit, Poverty, Firm default, Fraud, Unemployment, etc.
- Aim is to classify y, where y represents membership in a category
  - Qualitative, not necessarily ordered
  - We will focus for now in the binary case

The prediction question is, given a new X,
what is our best guess at the response category $\hat{y}$

Classification: Recap

$$1[p_i \geq c]$$

Agenda

1. Recap
2. Confusion Matrix
   - Accuracy
   - TNR
   - TNR
3. ROC curve
4. Imbalanced Classification
   - Metrics
   - Class rebalancing

Confusion Matrix

		$\hat{y}_i$ = 0	$\hat{y}_i$ = 1
y_i	0	TN	FP
	1	FN	TP

Accuracy

		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

$$\frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

TNR

		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

$$P[\hat{y} = 0|y = 0] = \frac{TN}{TN + FP} \tag{2}$$

TPR

		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

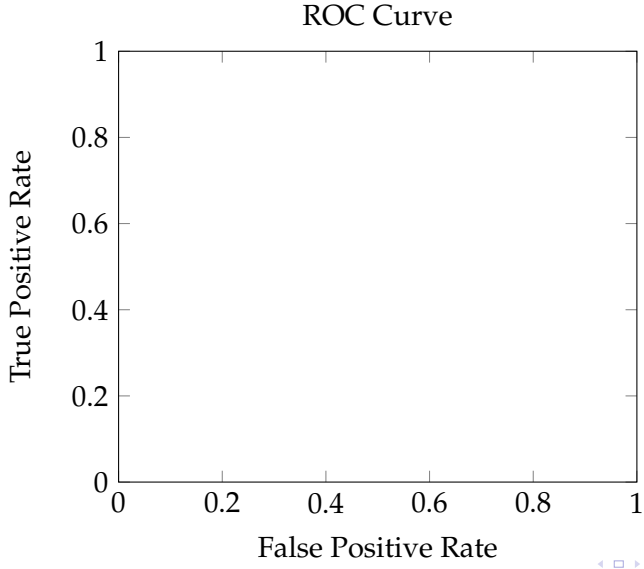
$$P[\hat{y} = 1 | y = 1] = \frac{TP}{TP + FN} \tag{3}$$

Agenda

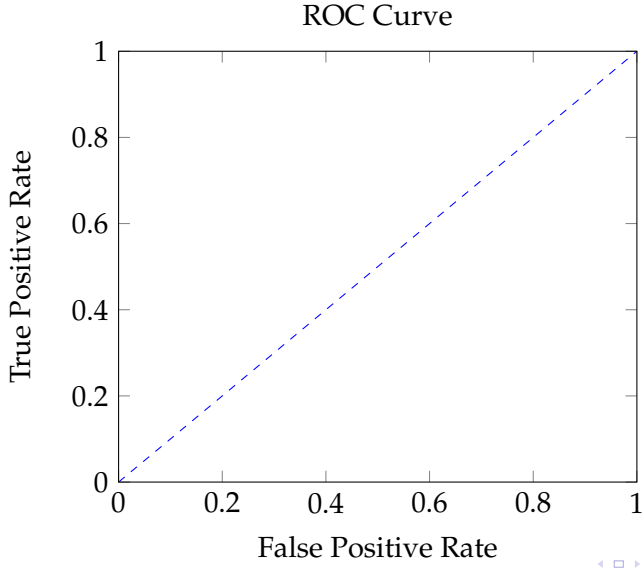
Trade-Off between Different Classification Thresholds

$$\hat{y}_i = 1[p_i \geq c]$$

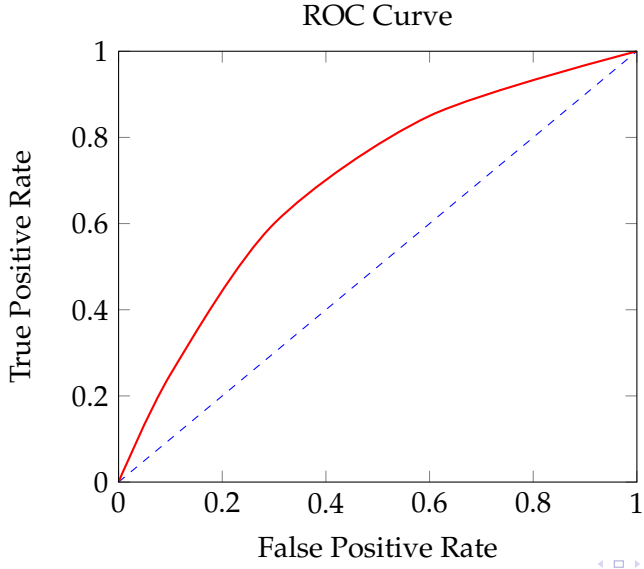
ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1

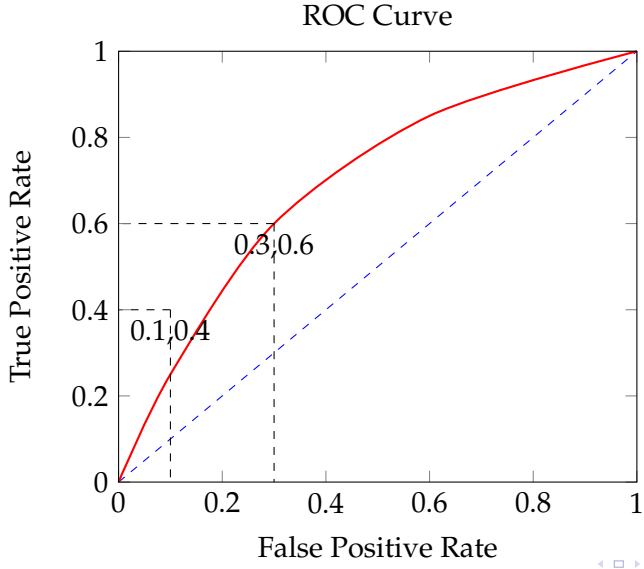
ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1

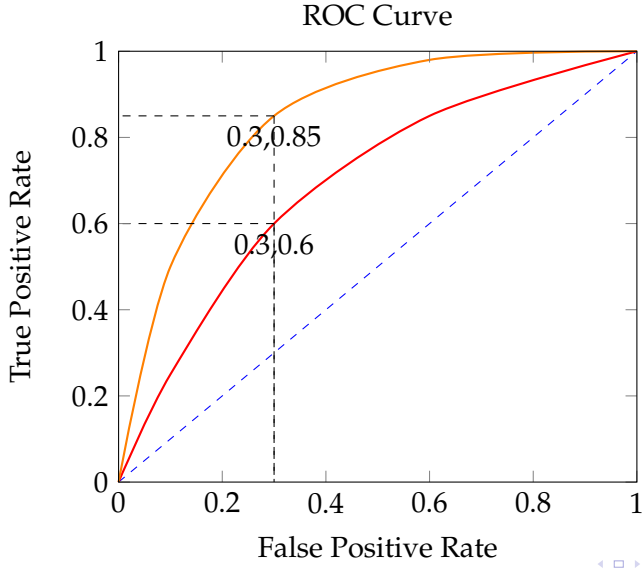
ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1

ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
0.3, 0.6
0.1, 0.4

ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
0.3,0.85
0.3,0.6

Example: Default

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

1. Recap
2. Confusion Matrix
   - Accuracy
   - TNR
   - TNR
3. ROC curve
4. Imbalanced Classification
   - Metrics
   - Class rebalancing

Imbalanced Classification: Motivation

- Interest in one of the classes: Poor, Default, Unemployed, Fraud
- Imbalanced classes pose a challenge

Imbalanced Classification: Motivation

- Interest in one of the classes: Poor, Default, Unemployed, Fraud
- Imbalanced classes pose a challenge

Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	$<$1% of the data set

Imbalanced Classification: Solutions

- Model Tuning
- Alternative Cutoffs
- Class rebalancing

TPR & PPV

		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

$$P[\hat{y} = 1|y = 1] = \frac{TP}{TP + FN} \tag{4}$$

TPR & PPV

		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

$$P[\hat{y} = 1|y = 1] = \frac{TP}{TP + FN} \tag{4}$$

$$P[y = 1|\hat{y} = 1] = \frac{TP}{TP + FP} \tag{5}$$

F-Scores

		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

F-Scores

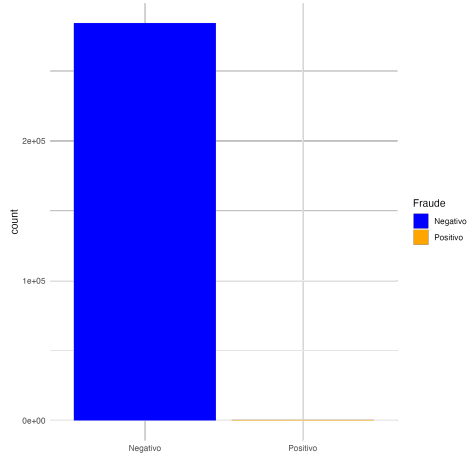
		$\hat{y}_i$	
		0	1
y_i	0	TN	FP
	1	FN	TP

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{(\beta^2 \times Precision + Recall)} \tag{7}$$

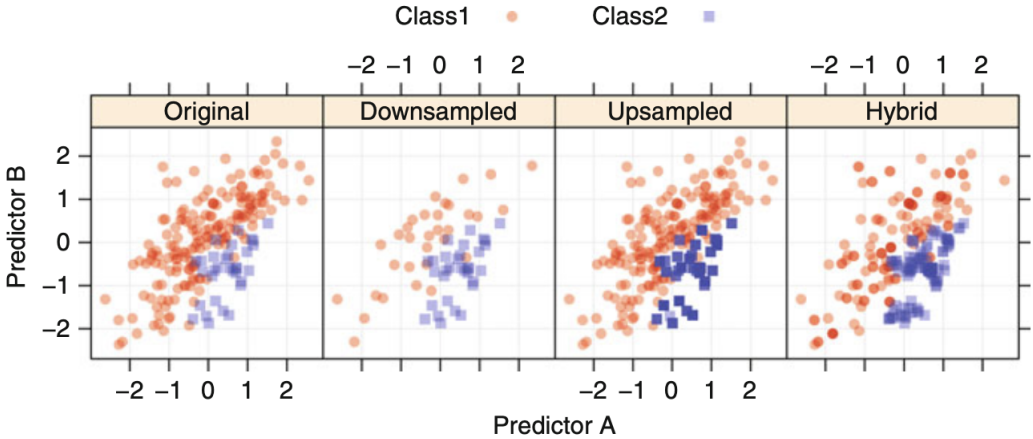
Example: Default

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Extreme Class Imbalance: Motivation

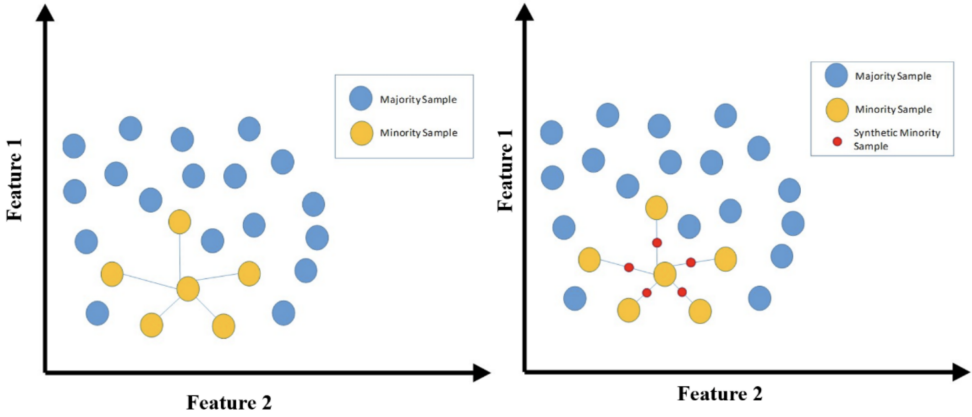


Class Rebalancing



Class Rebalancing: SMOTE

synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002)



Example: Fraud

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/