Data Mining and Machine Learning for Business

DSCI 5240

Evaluate Model Performance

Evaluating the performance of classification models

- Popular criteria
    - ~~Accuracy (misclassification)~~ rate: % of correct classifications
    - Confusion matrix
    - Lift curve/ROC curve
- Other evaluation criteria
    - Speed and scalability
    - Interpretability
    - Robustness

Accuracy (Misclassification) rate

- Accuracy rate $= \frac{\text{Number of correct classifications}}{\text{Number of instances in dataset}}$; = (9320 + 205)/10000 = 95.25%

- $Misclassification\ rate = 1\ - Accuracy\ Rate$ = (128 + 347)/10000 = 4.75%

		True default status		
		No	Yes	Total
Predicted default status	No	9320	128	9448
	Yes	347	205	552
	Total	9667	333	10000

Confusion Matrix

- A **confusion matrix** records the source of error:
  - **Type I error: False positives**
  - **Type II error: False negatives**

Predicted class (columns) / **Actual class** (rows)

	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

- Suppose 950 mails are sent out
- What is the accuracy rate?

Predicted class (columns) / **Actual class** (rows)

	Respond	Do not respond
Respond	250	40
Do not Respond	10	650

Confusion Matrix - Evaluation

- Below shows the performance of two classifiers. Which one is better based on accuracy?

Model 1 - Predicted class

		Respond	Do not respond
Actual class	Respond	5	5
	Do not Respond	40	950

- Accuracy = (5+950)/1000 = 95.5%
- Misclassification rate = 4.5%

Model 2 - Predicted class

		Respond	Do not respond
Actual class	Respond	10	0
	Do not Respond	90	900

- Accuracy= 91%?
- Misclassification rate = 9%?

Asymmetric costs of different types of errors

- Suppose cost of mailing to a non-responder is $1, and (net) lost revenue of not mailing to a responder is $20.
- Now from cost perspective, which classifier is better?

Model 1 - Predicted class

		Respond	Do not respond
Actual class	Respond	5	5
	Do not Respond	40	950

- Cost = 5*20 + 40*1 = $140

Model 2 - Predicted class

		Respond	Do not respond
Actual class	Respond	10	0
	Do not Respond	90	900

- Cost = 0 * 20 + 90 *1 = 90

The credit card default

Type - I

Type - II

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- What is Type I error rate? - 23 What is Type II error rate? - 252
- As a credit card company, which type of error would it like to avoid more? Type II error is more important.
- **Sensitivity**: the proportion of all positives that are correctly identified as positives - True positive rate – 81/333 =
- **Specificity**: the proportion of all negatives that are correctly identified as negatives – True negative rate 9644/9667 =

The credit card default

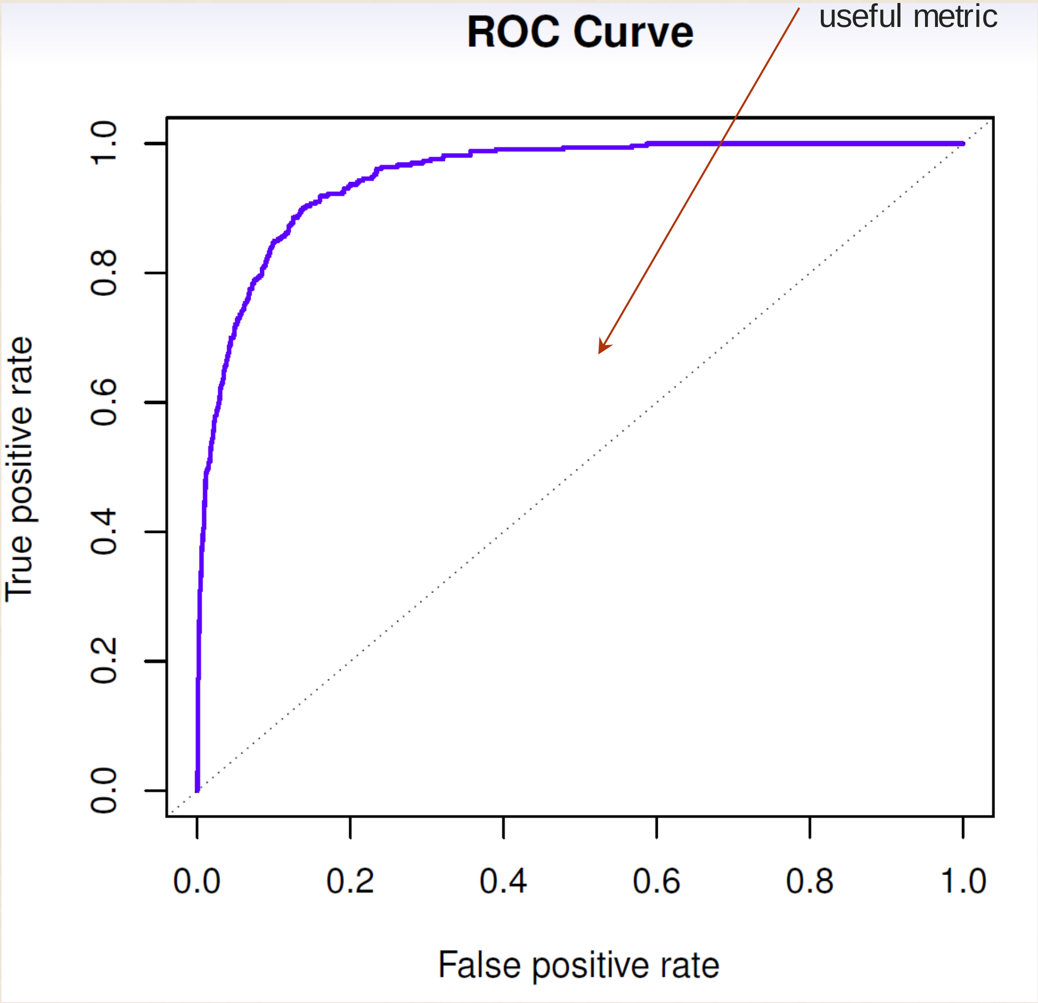
		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

- We adjust the threshold probability from 0.5 to 0.2
- Sensitivity increases
- It comes at a cost of decreasing specificity and slightly increasing error rate
- There is a trade-off between sensitivity and specificity

ROC curve

- ROC curve depicts the trade-off between **Sensitivity** vs **Specificity**
- It displays two types of errors for all possible thresholds
- False positive rate: **1 - specificity**
- The overall performance is given by the area under the curve: the larger the better
- An ideal ROC curve will hug the top left corner

Area under the ROC curve ("AUC") is a useful metric



Takeaways

- Discriminant analysis: models and assumptions
  - LDA
  - QDA
  - Naïve Bayes
- The comparison between them
- Evaluate performance of different methods
  - Accuracy rate
  - Confusion matrix
  - ROC curve