Mathematical Models

Module 1

The primary data science programming languages

- Python
- Perl
- R
- Matlab
- Java and C/C++
- Mathematica/Wolfram Alpha
- Excel

Standard Data Formats

The best computational data formats have several useful properties

- They are easy for computers to parse
- They are easy for people to read
- They are widely used by other tools and systems

The most important data formats/representations are:

- CSV (comma separated value) files
- XML (eXtensible Markup Language)
- SQL (structured query language) databases
- JSON (JavaScript Object Notation)
- Protocol buffers

A Taxonomy of Models

Linear vs. Non-Linear Models

Blackbox vs. Descriptive Models

First-Principle vs. Data-Driven Models

Stochastic vs. Deterministic Models

Flat vs. Hierarchical Models

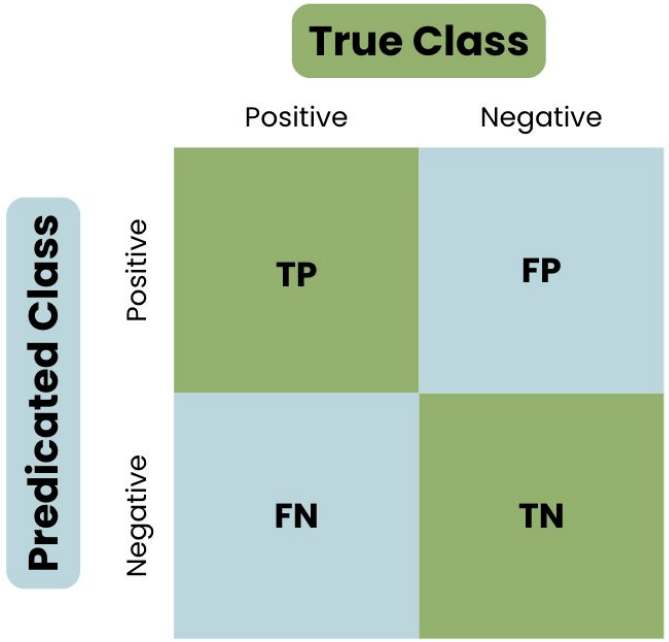
Evaluating Models

Evaluating Models

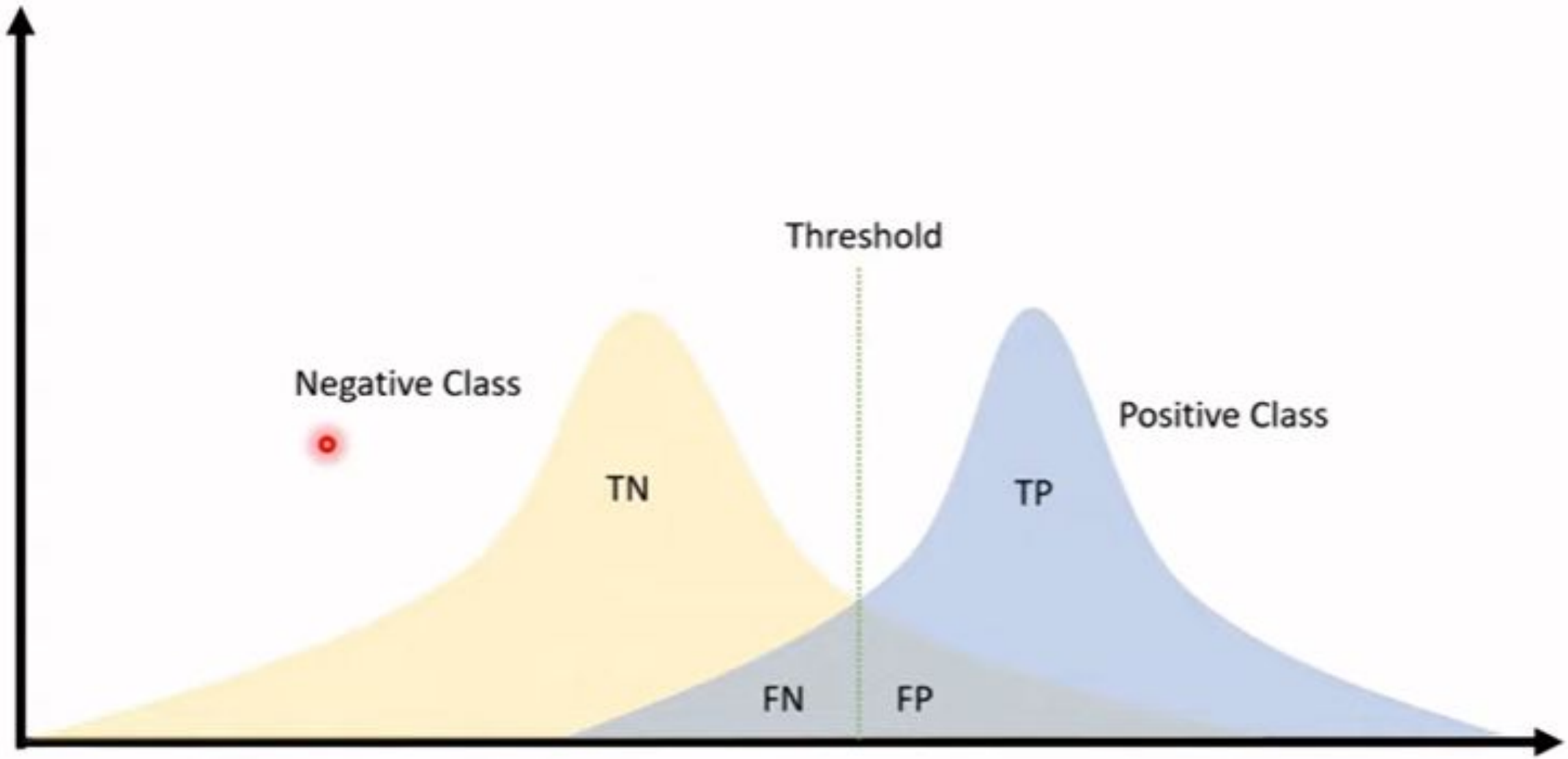
The confusion matrix for binary classifiers

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positives (TP)	False Negatives (FN)
	No	False Positives (FP)	True Negatives (TN)

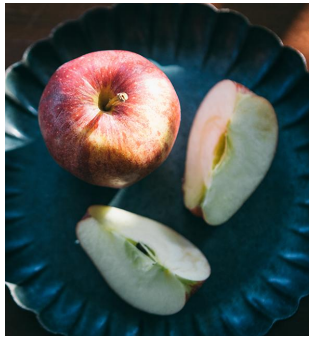
Confusion matrix or contingency table



- **True Positives (TP):** Here our classifier labels a positive item as positive, resulting in a win for the classifier.
- **True Negatives (TN)**: Here the classifier correctly determines that a member of the negative class deserves a negative label. Another win.
- **False Positives (FP):** The classifier mistakenly calls a negative item as a positive, resulting in a "**type I**" classification error.
- **False Negatives (FN):** The classifier mistakenly declares a positive item as negative, resulting in a **"type II"** classification error.

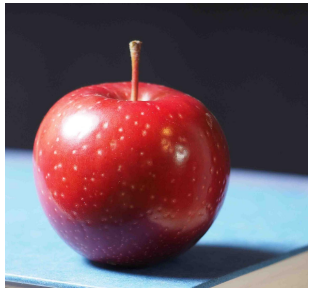

Threshold
Negative Class
Positive Class
TN
TP
FN
FP





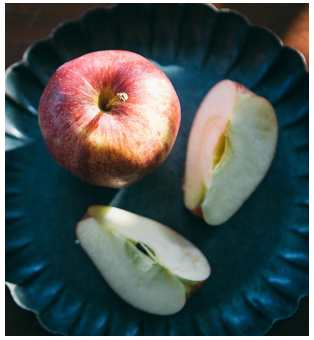







APPLE

NOT APPLE



APPLE

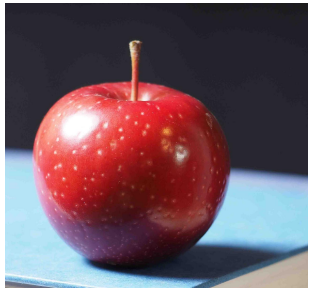
NOT APPLE

NOT APPLE

APPLE

NOT APPLE

APPLE



APPLE

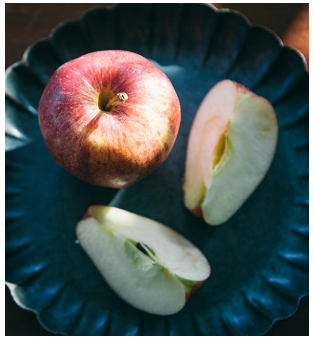
NOT APPLE

APPLE

APPLE

APPLE

NOT APPLE



APPLE

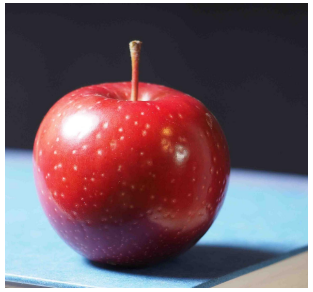
NOT APPLE

NOT APPLE

APPLE

NOT APPLE

APPLE



APPLE

NOT APPLE

APPLE

APPLE

Accuracy, Precision, Recall, and F-Score

- We must defend our classifier against two baseline opponents, the **sharp** and the **monkey**
- The **sharp** is the opponent who knows what evaluation system we are using and picks the baseline model which will do best according to it. The sharp will try to make the evaluation statistic look bad, by achieving a high score with a useless classifier. That might mean declaring all items positive, or perhaps all negative.
- In contrast, the **monkey** randomly guesses on each instance.
- To interpret our model's performance, it is important to establish by how much it beats both the sharp and the monkey.

Accuracy

- the ratio of the number of correct predictions over total predictions

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

$$accuracy\ (\%) = \frac{TP + TN}{TP + TN + FN + FP} \text{ x } 100$$

- Limitation

$$p = \frac{|positive|}{|positive| + |negative|} \ll \frac{1}{2}$$

Precision

- Precision measures how often this classifier is correct when it dares to say positive

$$precision = \frac{TP}{TP + FP}$$

Recall

- Recall measures how often you prove right on all positive instances

$$recall = \frac{TP}{TP + FN}$$

F-score (F1-score)

- The F-score (or sometimes F1-score) is such a combination, returning the harmonic mean of precision and recall

$$F = 2.\frac{precision.recall}{precision + recall}$$

	predicted class	
	yes	no
yes	$(pn)q$	$(pn)(1-q)$
no	$((1-p)n)q$	$((1-p)n)(1-q)$

	predicted class	
	yes	no
yes	$(pn)q$	$(pn)(1-q)$
no	$((1-p)n)(1-q)$	$((1-p)n)q$

Figure 7.4: The expected performance of a monkey classifier on n instances, where $p \cdot n$ are positive and $(1-p) \cdot n$ are negative. The monkey guesses positive with probability q (left). Also, the expected performance of a balanced classifier, which somehow correctly classifies members of each class with probability q (right).

	Monkey		Sharp		Balanced Classifier				
q	0.05	0.5	0.0	1.0	0.5	0.75	0.9	0.99	1.0
accuracy	0.905	0.5	0.95	0.05	0.5	0.75	0.9	0.99	1.
precision	0.05	0.05	—	0.05	0.05	0.136	0.321	0.839	1.
recall	0.05	0.5	0.	1.	0.5	0.75	0.9	0.99	1.
F score	0.05	0.091	—	0.095	0.091	0.231	0.474	0.908	1.

Accuracy, Precision, Recall, and F-Score

- Accuracy is a misleading statistic when the class sizes are substantially different
- Recall equals accuracy if and only if the classifiers are balanced
- High precision is very hard to achieve in unbalanced class sizes
- F-score does the best job of any single statistic, but all four work together to describe the performance of a classifier

Receiver-Operator Characteristic (ROC) Curves

ROC

- Sensitivity: True positive rate (recall)

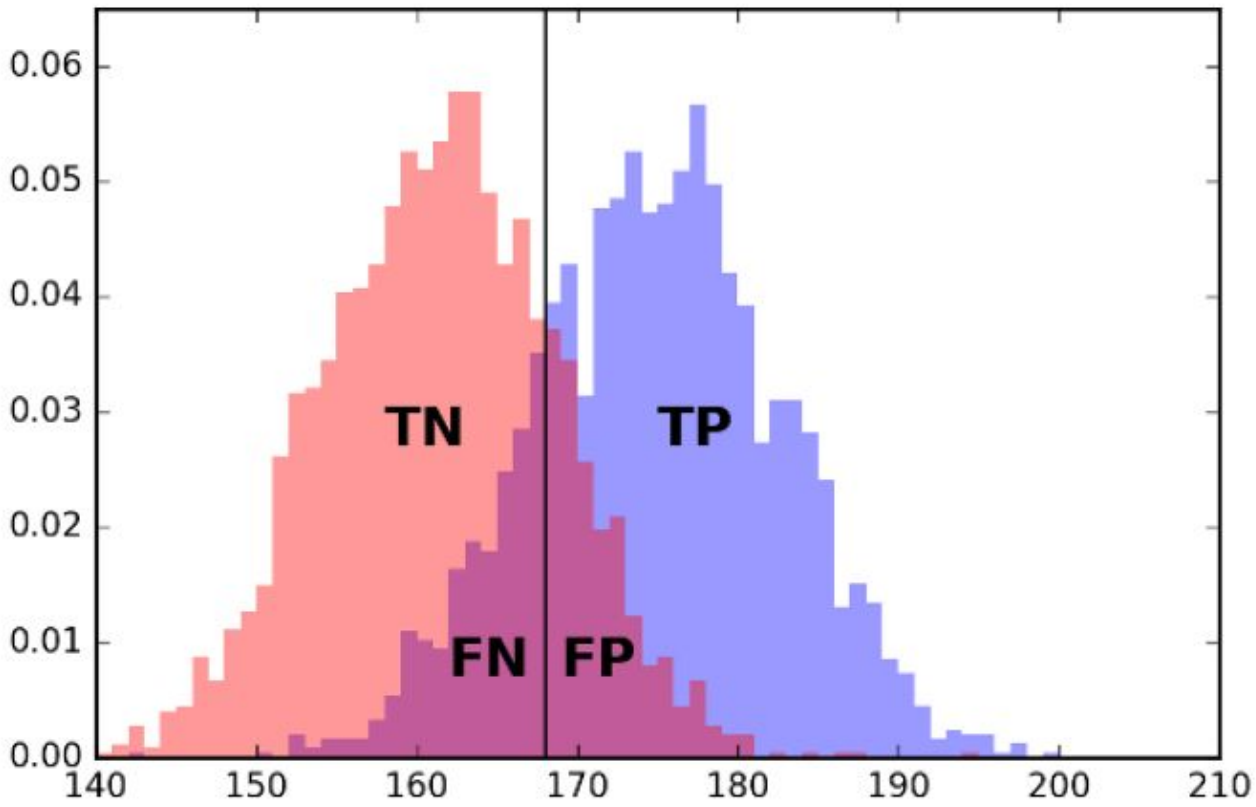
$$recall = sensitivity = \frac{TP}{TP + FN}$$

- Specificity: True negative rate

$$specificity = \frac{TN}{TN + FP}$$

$$False\ positive\ Rate = 1 - specificity$$

$$specificity = \frac{FP}{TN + FP}$$



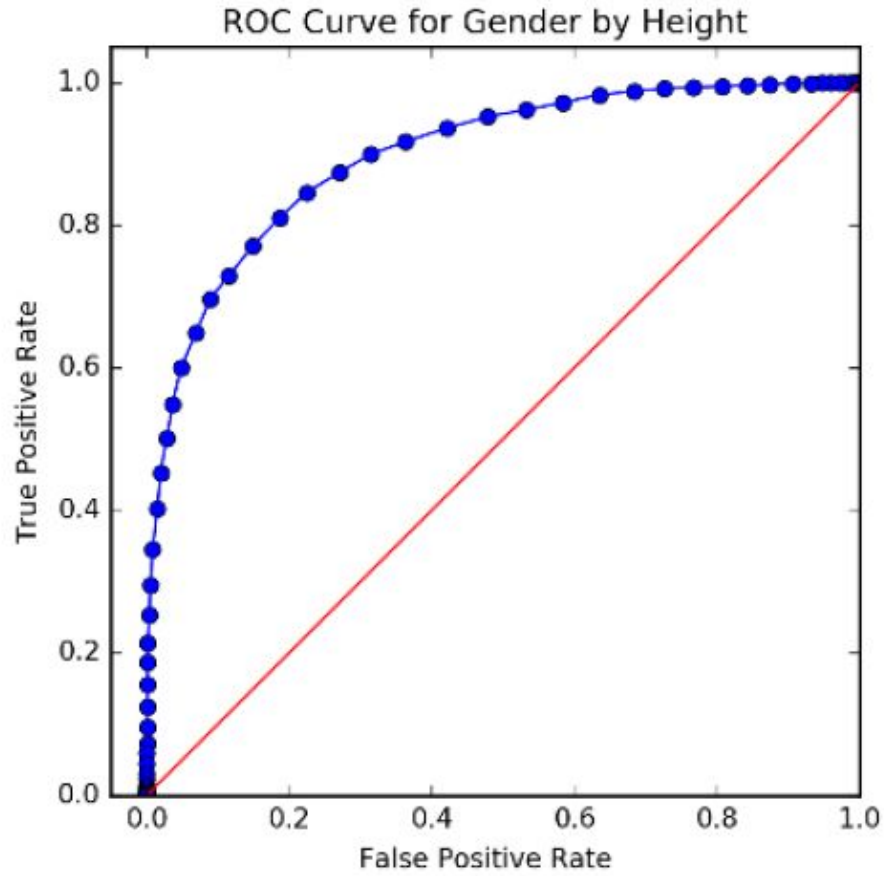
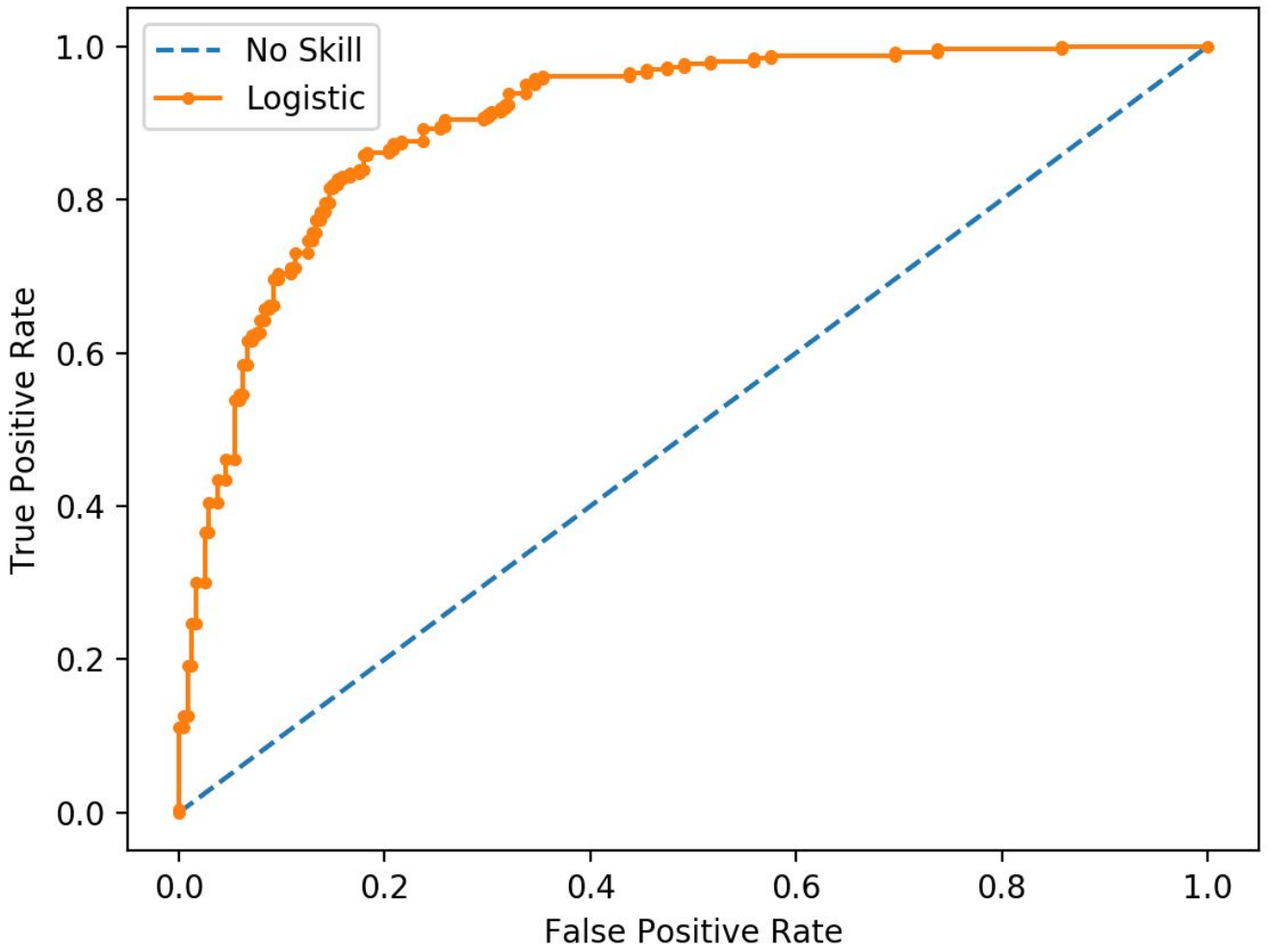


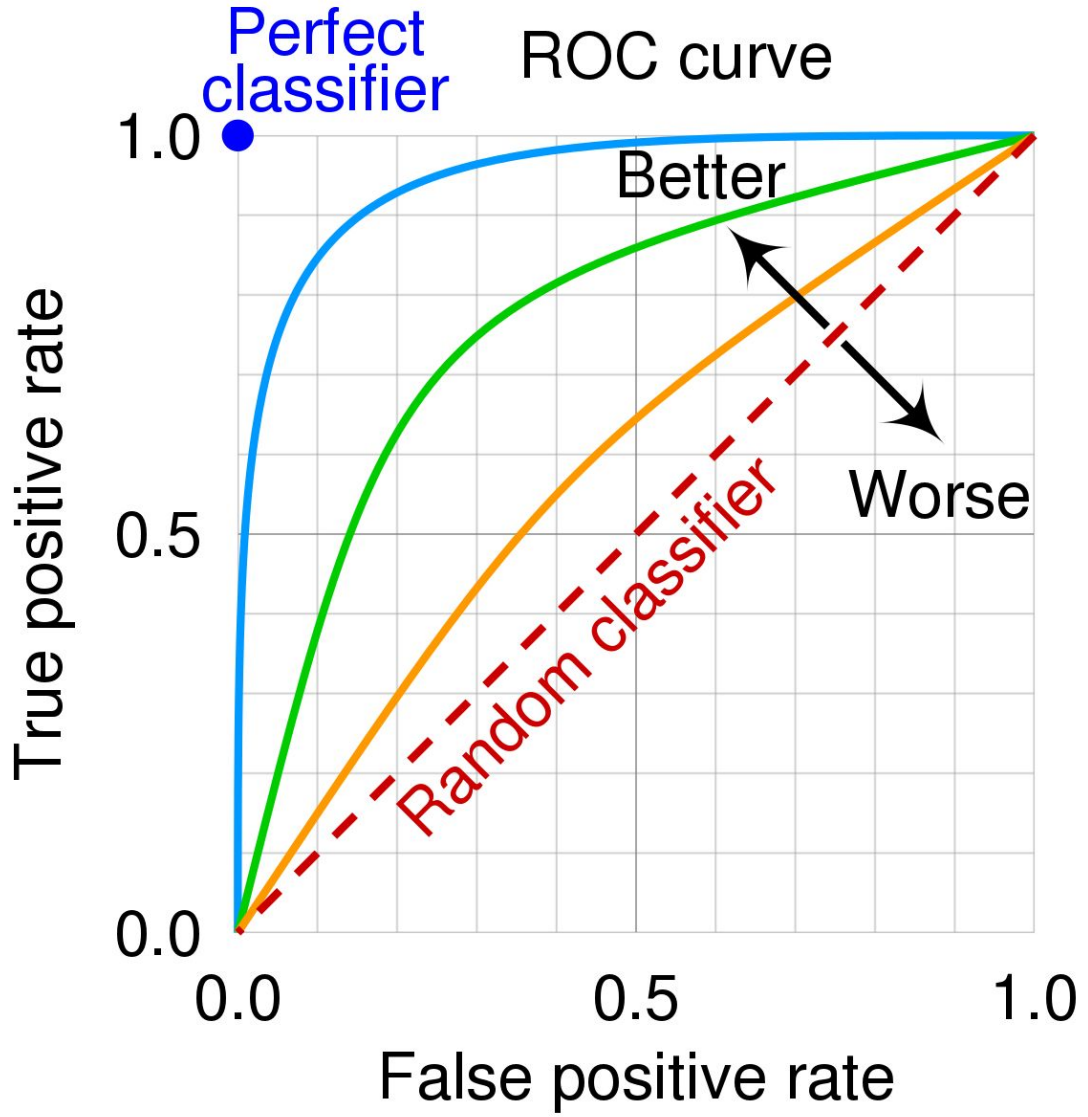
Figure 7.6: The ROC curve helps us select the best threshold to use in a classifier, by displaying the trade-off between true positives and false positive at every possible setting. The monkey ROCs the main diagonal here.

ROC curve – logistic regression



AUC (Area under the ROC curve)

- The area under the ROC curve (AUC) is often used as a statistic measuring the quality of scoring function defining the classifier.



Evaluating Multiclass Systems

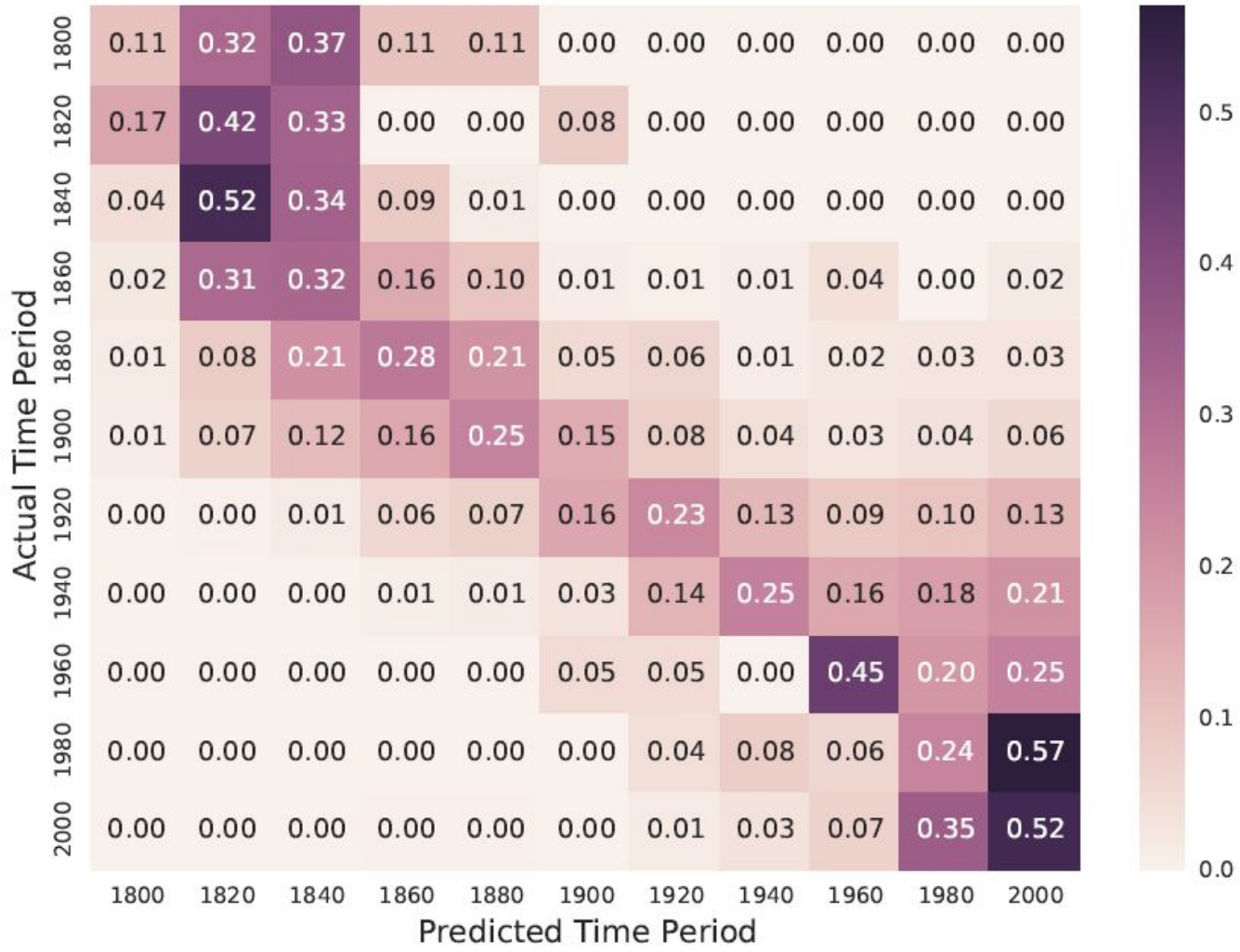


Figure 7.7: Confusion matrix for a document dating system: the main diagonal reflects accurate classification.

$$precision_i = C[i,i] / \sum_{j=1}^{d} C[j,i].$$

$$recall_i = C[i,i] / \sum_{j=1}^{a} C[i,j].$$

Evaluating Value Prediction Models

- Error Statistics
  - error is a function of the difference between a forecast y' =f(x) and the actual result y
- Absolute error :

$$\Delta = y' - y$$

- Relative error:

$$\varepsilon = (y - y')/y$$

- Squared error

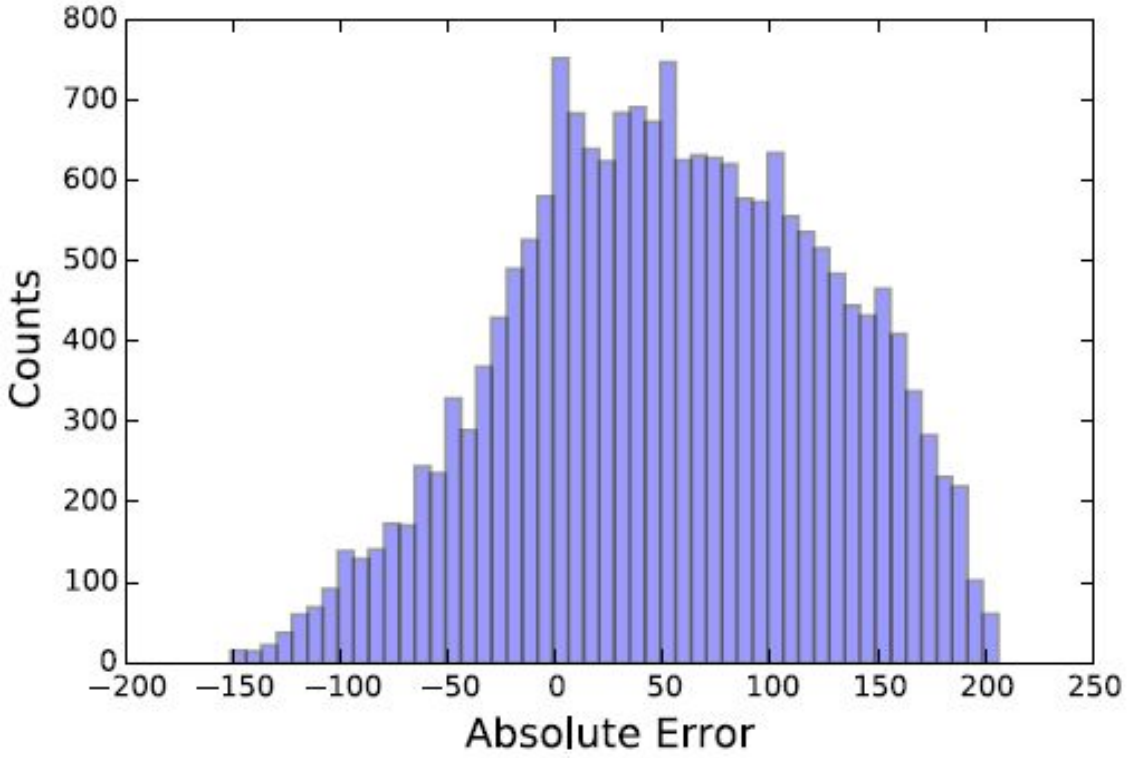
$$\Delta^2 = (y' - y)^2$$

- Mean squared error (MSE):

$$MES(y, y') = \frac{1}{n}\sum_{i=1}^{n}(y' - y)^2$$

- Root mean squared (RMSD)

$$RMSD(\theta) = \sqrt{MES(y, y')}$$



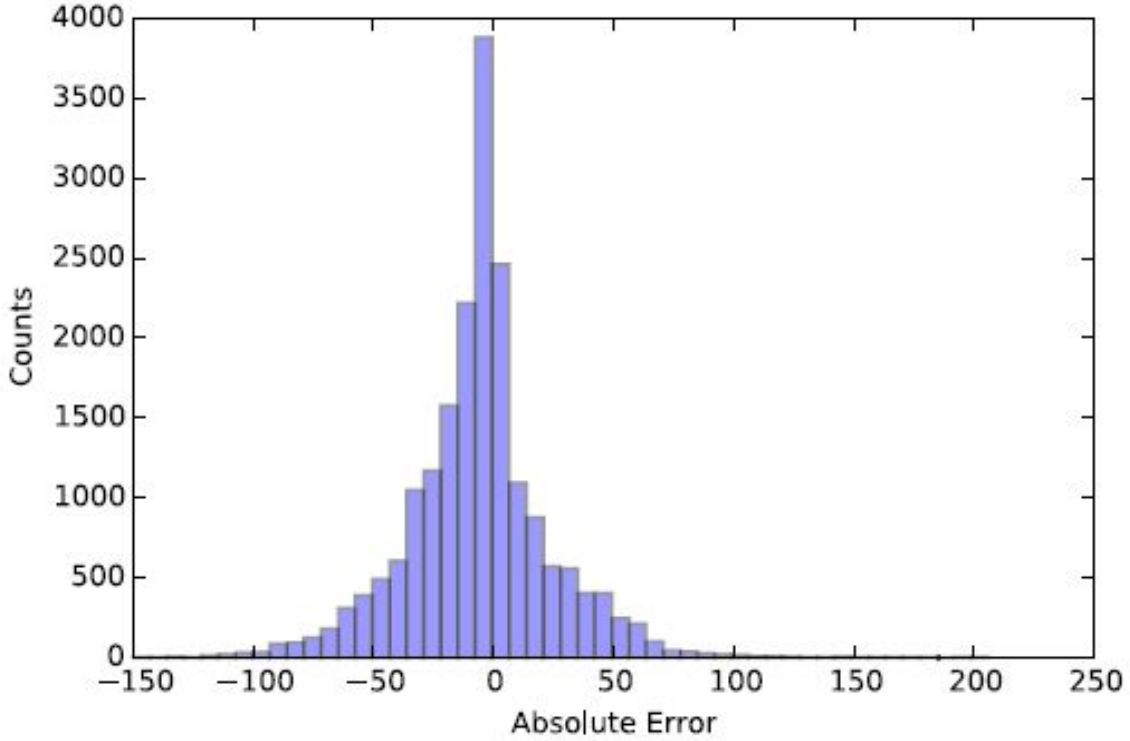


Figure 7.8: Error distribution histograms for random (left) and naive Bayes classifiers predicting the year of authorship for documents (right).

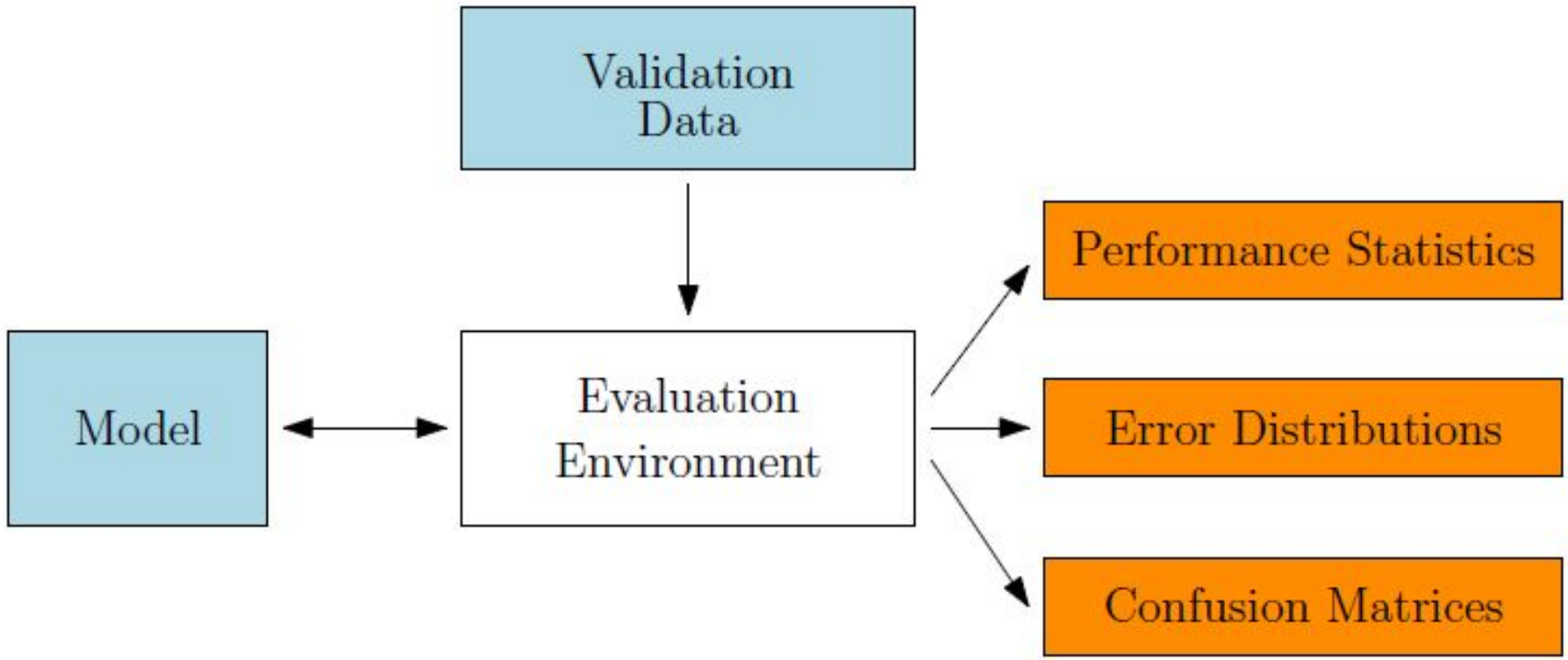


Figure 7.9: Block diagram of a basic model evaluation environment.