Predicted

Actual

Classes	A	B	C	D	Total
A	50	10	30	20	110
B	15	20	30	15	80
C	20	10	30	40	100
D	15	15	30	50	110
Total	100	55	120	125	800

6. Performance Evaluation

Model Selection

As a consequence of the **No-Free-Lunch Theorem**, no single classifier can be the best performer for every problem.

Model selection must take into account:

- the **nature** of the available data
- the **relative frequencies of the classification sub-groups**
- the **stated classification goals**
- how easily the model lends itself to **interpretation** and **statistical analysis**
- how much **data preparation** is required

Model Selection

Model selection must take into account (continued):

- whether it can accommodate various data types and missing observations
- whether it performs well with large datasets, and
- whether it is **robust** against small data departures from theoretical assumptions.

Past success is not a guarantee of future success – it is the analyst's responsibility to try a **variety of models**.

But how can the "**best**" model be selected?

Classification Errors

When attempting to determine what kind of music a new customer would prefer, there is no real **cost** in making a mistake; if, on the other hand, the classifier attempts to determine the presence or absence of cancerous cells in lung tissue, mistakes are **more consequential**.

Several metrics can be used to assess a classifier's performance, depending on the context.

Binary classifiers are simpler and have been studied far longer than multi-level classifiers; consequently, a larger body of evaluation metrics is available for these classifiers.

Binary Classifiers

		Predicted: Category I	Predicted: Category II	Total
Actuals	Category I	TP	FN	AP
	Category II	FP	TN	AN
	Total	PP	PN	T

TP, TN, FP, FN: **True Positives**, **True Negatives**, **False Positives**, and **False Negatives**, respectively.

Perfect classifiers would have $FP, FN = 0$, but that rarely ever happens in practice (and not ideal, in a way).

Metrics:

- sensitivity = $TP/(TP + FN)$
- specificity = $TN/(FP + TN)$
- precision = $TP/(TP + FP)$
- recall = $TP/(TP + FN)$
- negative predictive value = $TN/(TN + FN)$
- false positive rate = $FP/(FP + TN)$
- false discovery rate = $FP/(FP + TP)$
- false negative rate = $FP/(FN + TP)$
- accuracy = $(TP + TN)/T$

Other metrics:

F_1-score, ROC AUC, informedness, markedness, Matthews' Correlation Coefficient (MCC), etc.

Session 2

		Predicted A	Predicted B	Total	
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

		Predicted A	Predicted B	Total	
Actuals	A	54	0	54	66.7%
	B	16	11	27	33.3%
Total		70	11	81	
		86.4%	13.6%		

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

Both classifiers have an accuracy of 80%; the second classifier makes some wrong predictions for A, but never for B; the first classifier makes mistakes for both classes. The second classifier mistakenly predicts occurrence A as B on 16 occasions, but the first one only does so 6 times. Which one is best depends on the **cost of misclassification**.

Multi-Level Classifiers

It is preferable to select metrics that generalize more readily to **multi-level classifiers.**

Accuracy: proportion of correct predictions amid all the observations

- value ranges from 0% to 100%
- the higher the accuracy, the better the match
- a predictive model with high accuracy may be useless thanks to the **Accuracy Paradox**

Matthews Correlation Coefficient (MCC): useful even when the classes are of very different sizes

- correlation coefficient between actual and predicted classifications
- range varies from -1 to 1
- if $MCC = 1$, predicted and actual responses are identical
- if $MCC = 0$, the classifier performs no better than a random prediction (“flip of a coin”).

Multi-Level Classifiers

MCC: 69.7%
Accuracy: 78.3%
Pearson: 0.13161
Hist: 30.0%

		Predicted							
		Maltreatment			Risk				
		Unfounded	Suspected	Substantiated	No	Yes	Unknown	*Total*	
Actuals: Maltreatment	Unfounded	4,577	-	-	198	6	-	*4,781*	*29.2%*
	Suspected	-	965	-	29	2	-	*995*	*6.1%*
	Substantiated	-	-	6,187	116	35	2	*6,339*	*38.7%*
Actuals: Risk	No	894	-	763	949	19	9	*2,632*	*16.1%*
	Yes	123	-	520	122	111	5	*880*	*5.4%*
	Unknown	212	-	303	184	21	24	*745*	*4.6%*
Total		*5,805*	*965*	*7,772*	*1,597*	*194*	*40*	***16,372***	
		35.5%	*5.9%*	*47.5%*	*9.8%*	*1.2%*	*0.2%*		

Regression Performance Evaluation

For numerical targets y with predictions $\hat{y}$, metrics include:

- **mean squared** and **mean absolute errors**

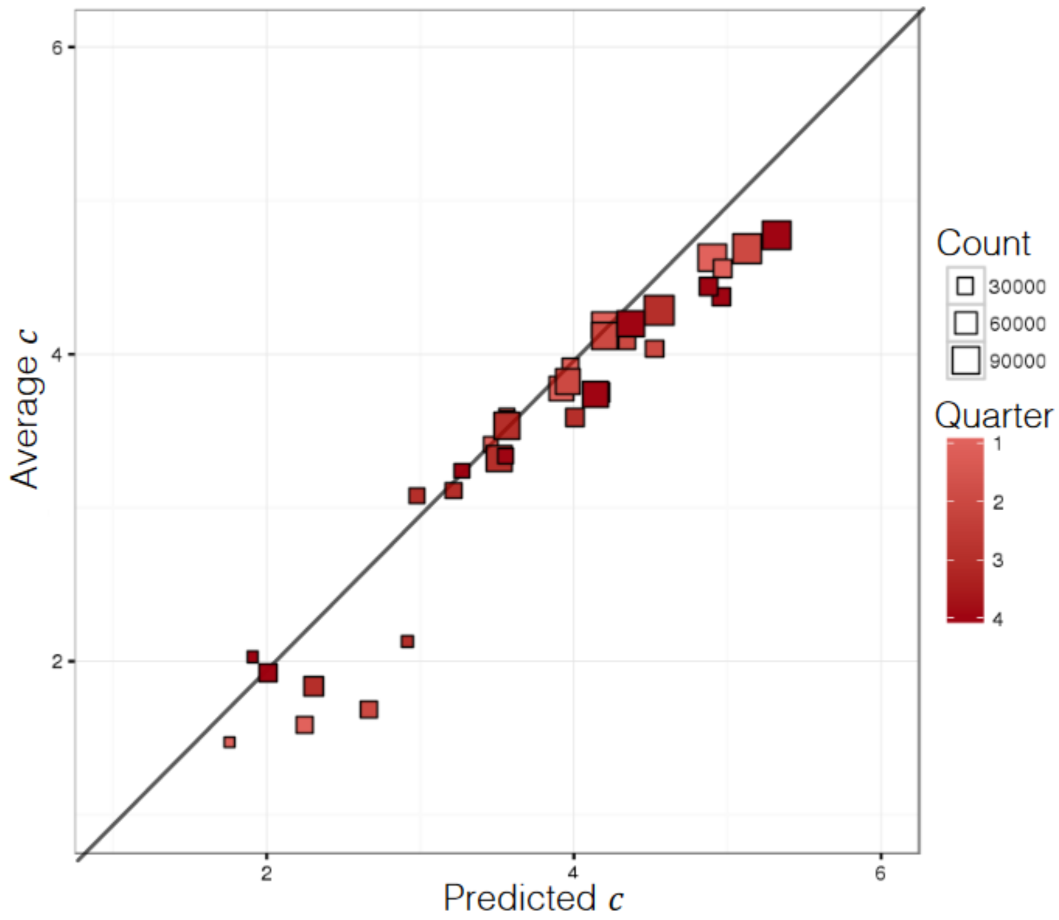
$$\mathrm{MSE} = \mathrm{mean}\{(\hat{y}_i - y_i)^2\}, \ \mathrm{MAE} = \mathrm{mean}\{|\hat{y}_i - y_i|\}$$

- **normalized mean squared** and **normalized mean absolute errors**

$$\mathrm{NMSE} = \frac{\mathrm{mean}\{(\hat{y}_i - y_i)^2\}}{\mathrm{mean}\{(\bar{y} - y_i)^2\}}, \mathrm{NMAE} = \frac{\mathrm{mean}\{|\hat{y}_i - y_i|\}}{\mathrm{mean}\{|\bar{y} - y_i|\}}$$

- **mean average percentage error** $\mathrm{MAPE} = \mathrm{mean}\left\{\frac{|\hat{y}_i - y_i|}{y_i}\right\}$
- **correlation** $\rho_{\hat{y},y}$

Regression Performance Evaluation



Suggested Reading

Performance Evaluation

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Classification and Value Estimation

- [Performance Evaluation]

Regression and Value Estimation

*Statistical Learning (advanced)

- [Model Evaluation]

Exercises

Performance Evaluation

We continue the UniversalBank example. The confusion matrices for the predictions of trees A and B on the remaining 2000 testing observations are shown below.

1. Using the appropriate matrices, compute the performance evaluation metrics for each of the trees (on the testing set).

2. If customers who would not accept a personal loan get irritated when offered a personal loan, what tree should *the* marketing group use to maintain good customer relations?

Tree A

		Predicted A	Predicted B	Total	
Actuals	A	1792	19	1811	90.55%
	B	18	171	189	9.45%
Total		1810	190	2000	
		90.50%	9.50%		

Tree B

		Predicted A	Predicted B	Total	
Actuals	A	1801	10	1811	90.55%
	B	64	125	189	9.45%
Total		1865	135	2000	
		93.25%	6.75%		