Chapter 4. Measuring Classifier Performance

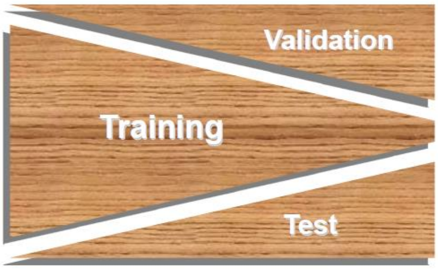
4.1 Honest Assessment

The Optimism Principle

Evaluating the performance of a classifier on the same data used to train the classifier usually leads to an optimistically biased assessment. The model can be overly sensitive to peculiarities of the particular training data, in addition to true features of their joint distribution. This is called *overfitting*.

The more flexible the underlying model and the less plentiful the data, the more overfitting becomes a problem. Large differences between the performance on the training and test data sets usually indicate overfitting.

Data Splitting



- Comparison
- Selection
- Tuning

- Final Assessment
- one-fourth to one-half of the development data

The simplest strategy for correcting the optimism bias is to hold out a portion of the development data for assessment. The model is fit to the remainder (*training data set*) and performance is evaluated on the holdout portion (*test data set*). After assessment, it is common practice to refit the final model on the entire undivided data set.

When the holdout data are used for comparing, selecting, and tuning models, the holdout sample is more correctly called a *validation data set*, not a test set. The test set is used for a final assessment of a fully specified classifier. If model tuning and a final assessment are both needed, then the data should be split three ways into training, validation, and test sets. In some applications, the test set is gathered from a different time or location.

Other Approaches

Data splitting is a simple but costly technique. When data are scarce, it is inefficient to use only a portion for training. Furthermore, when the test set is small, the performance measures might be unreliable because of high variability. For small and moderate data sets, *v*-fold cross validation is a better strategy.

Another approach that is frugal with the data is to assess the model on the same data set that was used for training but to penalize the assessment for optimism. The appropriate penalty can be determined theoretically or by using computationally intensive methods such as bootstrap.

4.2 Misclassification

Confusion Matrix

		Predicted Class		
		0	1	
Actual Class	0	True Negative	False Positive	Actual Negative
	1	False Negative	True Positive	Actual Positive
		Predicted Negative	Predicted Positive	

The fundamental assessment tool is the confusion matrix. The *confusion matrix* is a crosstabulation of the actual and predicted classes. If quantifies the confusion of the classifier. The event of interest, whether it in unfavorable or favorable, is often called a positive, although this convention is arbitrary. The simplest performance statistics are *accuracy* and *error rate*.

$$accuracy = \frac{true\ positives\ and\ negatives}{total\ cases}$$

$$error\ rate = \frac{false\ positives\ and\ negatives}{total\ cases}$$

Sensitivity and Positive Predicted Value

Two specialized measures of classifier performance are *sensitivity* and *positive predicted value*. The analogs to these measures for true negatives are *specificity* and *negative predicted value*.

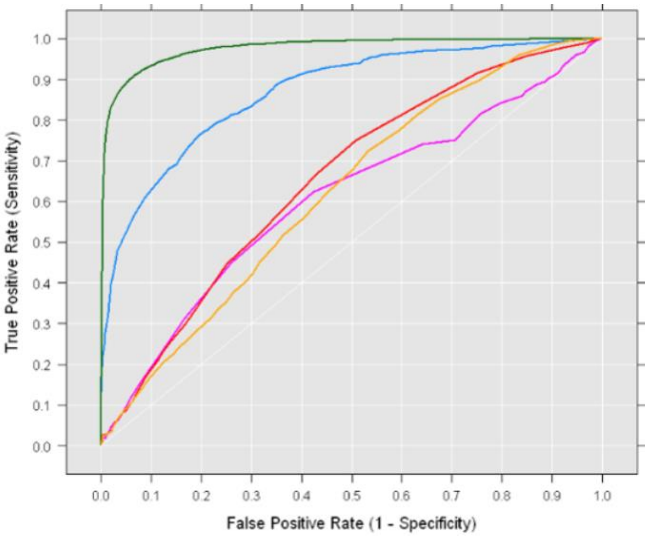
$$sensitivity = \frac{TP}{TP + FN}$$

$$positive\ predicted\ value\ (PV+) = \frac{TP}{TP + FP}$$

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

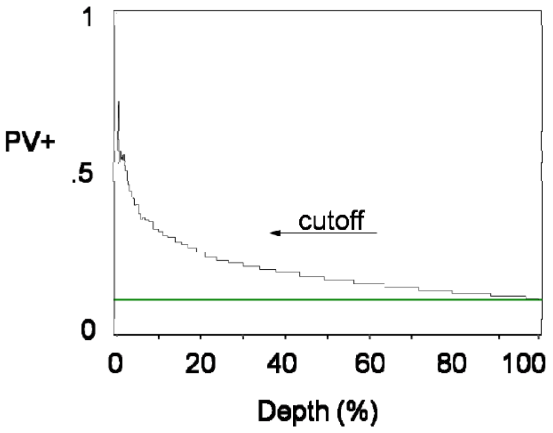
$$negatvie\ predicted\ value\ (PV-) = \frac{TN}{TN + FN}$$

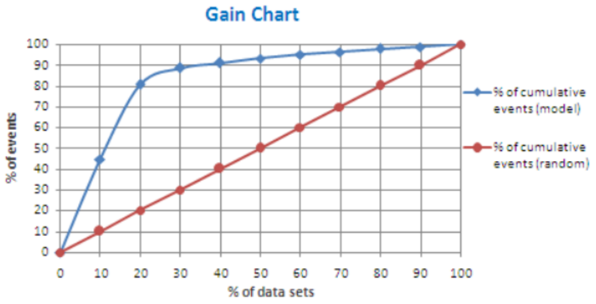
ROC Curve

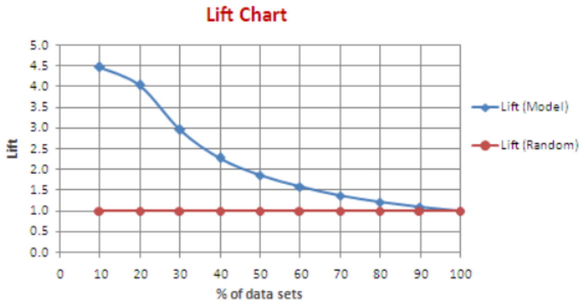


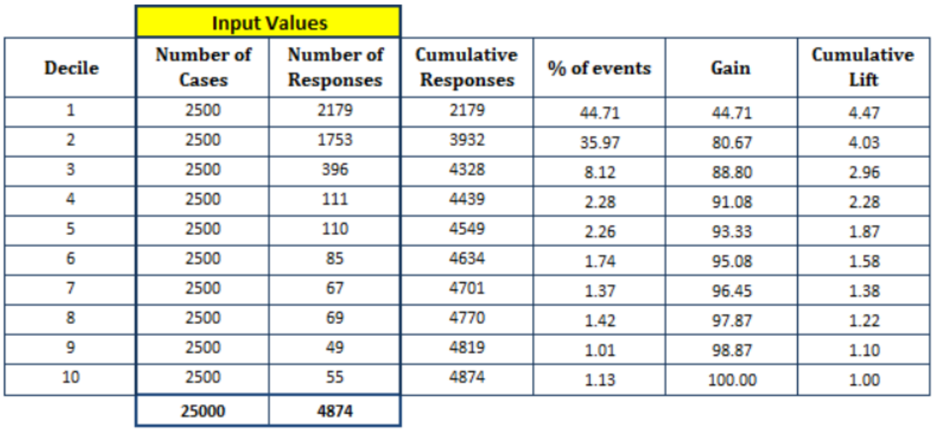
The receiver operating characteristic (ROC) curve displays the sensitivity and specificity for the entire range of cutoff values. As the cutoff decreases, more and more cases are allocated to class 1. Hence, the sensitivity increases and specificity decreases. As the cutoff increases, more and more cases are allocated to class 0. Hence, the sensitivity decreases and specificity increases. Consequently, the ROC curve intersects (0,0) and (1,1).

Gains Chart







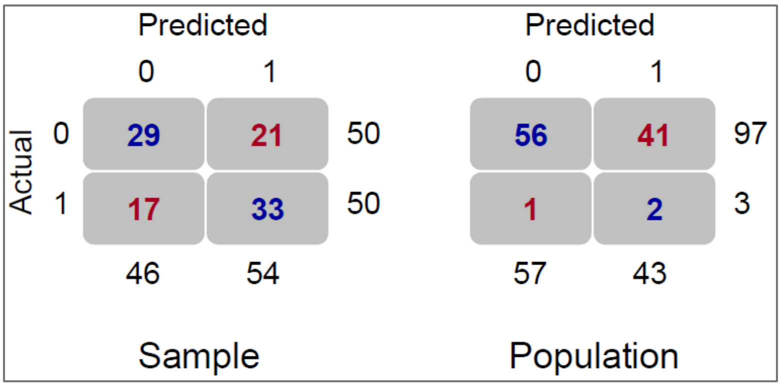


	Input Values					
Decile	Number of Cases	Number of Responses	Cumulative Responses	% of events	Gain	Cumulative Lift
1	2500	2179	2179	44.71	44.71	4.47
2	2500	1753	3932	35.97	80.67	4.03
3	2500	396	4328	8.12	88.80	2.96
4	2500	111	4439	2.28	91.08	2.28
5	2500	110	4549	2.26	93.33	1.87
6	2500	85	4634	1.74	95.08	1.58
7	2500	67	4701	1.37	96.45	1.38
8	2500	69	4770	1.42	97.87	1.22
9	2500	49	4819	1.01	98.87	1.10
10	2500	55	4874	1.13	100.00	1.00
	25000	4874				

The *depth* of a classification rule is the total proportion of cases that were allocated to class 1. The (cumulative) *gains chart* displays the positive predicted value and depth for a range of cutoff values. As the cutoff decreases, more and more cases are allocated to class 1; hence, the depth increases and the PV+ approaches the marginal event rate. When the cutoff is minimum, then 100% of the cases are selected and the response rate is ρ_1. As the cutoff increases the depth decreases. A model with good predictive power would have increasing PV+ (response rate) as the depth decreases. If the posterior probabilities were arbitrarily assigned to the cases, then the gains chart would be a horizontal line at ρ_1.

A plot of sensitivity versus depth is sometimes called a *Lorentz curve*, *concentration curve*, or a *lift curve*. The *lift* is PV+/ρ_1, so for a given depth, there are lift × *more responders targeted by the model than by random chance*.

Oversampled Test Set



Sample

		Predicted 0	Predicted 1	
Actual	0	29	21	50
Actual	1	17	33	50
		46	54	

Population

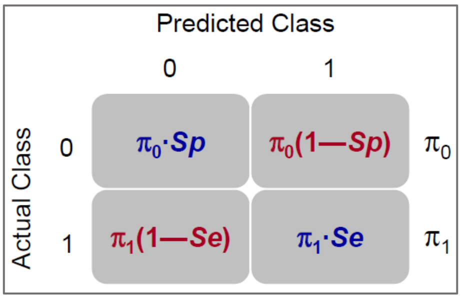
		Predicted 0	Predicted 1	
Actual	0	56	41	97
Actual	1	1	2	3
		57	43	

$$sensitivity = \frac{33}{33+17} = 0.66 = \frac{2}{1+2}$$

$$specificity = \frac{29}{29+21} = 0.58 = \frac{56}{56+41}$$

If the holdout data were obtained by splitting oversampled data, then they are oversampled as well. If the proper adjustments were made when the model was fitted, then the predicted posterior probabilities are correct. However, the confusion matrices would be incorrect because the event cases are over-represented. Sensitivity and specificity, however, are not affected by separate sampling because they do not depend on the proportion of each class in the sample.

Adjustments for Oversampling



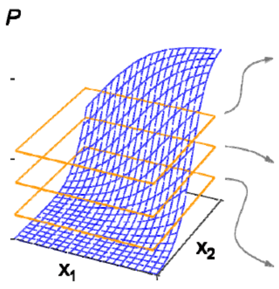
		Predicted Class 0	Predicted Class 1	
Actual Class	0	$\pi_0 \cdot Sp$	$\pi_0(1—Sp)$	π_0
Actual Class	1	$\pi_1(1—Se)$	$\pi_1 \cdot Se$	π_1

Knowing sensitivity, specificity, and the priors is sufficient for adjusting the confusion matrix for oversampling. Note that these adjustments are equivalent to multiplying the cell counts by their sample weights, for example

$$\mathrm{TP}_{\mathrm{sample}} \cdot wt_1 = \mathrm{TP}_{\mathrm{sample}} \frac{\pi_1}{\rho_1} = \mathrm{TP}_{\mathrm{sample}} \cdot \pi_1 \frac{n}{\mathrm{Tot\ Pos}_{\mathrm{sample}}} = n \cdot \pi_1 \cdot Se$$

4.3 Allocation Rules

Cutoffs



0	1	Se	Sp
70	5	.64	.93
9	16		
66	9	.84	.88
4	21		
57	18	.96	.76
1	24		

Different cutoffs produce different allocations and different confusion matrices. To determine the optimal cutoff, a performance criterion needs to be defined. For realistic data, there is a trade-off between sensitivity and specificity. Higher cutoffs decrease sensitivity and increase specificity. Lower cutoffs decrease specificity and increase sensitivity.

Profit Matrix

A formal approach to determining the optimal cutoff uses statistical decision theory. The decision-theoretic approach starts by assigning profit margins to true positives and loss margins to false positives. The optimal decision rule maximizes the total expected profit.

A *profit matrix* is used to assign costs to undesirable outcomes and profits to desirable outcomes.

		Decision	
		0	1
Actual Class	0	δ_{TN}	δ_{FP}
	1	δ_{FN}	δ_{TP}

Bayes Rule:

Decision 1 if

$$P > \frac{1}{1+\left(\frac{\delta_{\mathrm{TP}}-\delta_{\mathrm{FN}}}{\delta_{\mathrm{TN}}-\delta_{\mathrm{FP}}}\right)}$$

The *Bayes rule* is the decision rule that maximizes the expected profit. In the two-class situation, the *Bayes rule* can be determined analytically. Using the symbols in the above profit matrix, it a customer is solicited then the expected profit is

$$p(\delta_{TP}) + (1-p)(\delta_{FP})$$

where *p* is the true posterior probability that a case belongs to class 1. If a customer is not solicited, then the expected profit is

$$p(\delta_{FN}) + (1-p)(\delta_{TN})$$

Therefore, the optimal rule allocates a case to class 1 if

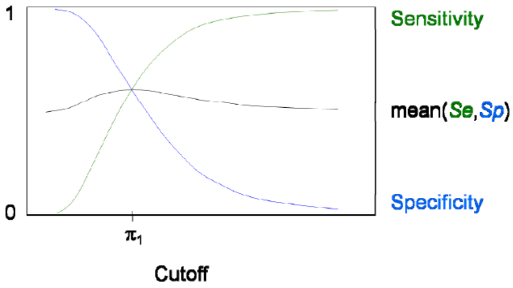
$$p(\delta_{TP}) + (1-p)(\delta_{FP}) > p(\delta_{FN}) + (1-p)(\delta_{TN)}$$

otherwise allocate the case to class 0. Solving for *p* gives the optimal cutoff probability. Because *p* must be estimated from the data, the *plug-in Bayes rule* is used in practice.

$$\hat{p} > \frac{1}{1+(\frac{\delta_{TP}-\delta_{FN}}{\delta_{TN}-\delta_{FP}})}$$

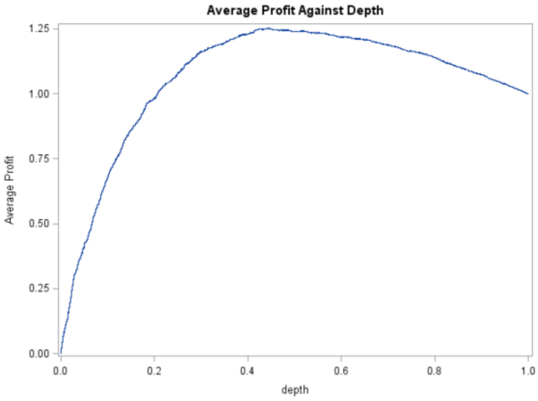
Consequently, the *plug-in Bayes rule* might not achieve the maximum profit if the estimate of the posterior probability is poorly estimated.

Classifier Performance



In many situations, gathering profit information can be difficult. One recommendation is to use a cutoff of π_1. The central cutoff, π_1, tends to maximize the mean of sensitivity and specificity. Because increasing sensitivity usually corresponds to decreasing specificity, the central cutoff tends to equalize sensitivity and specificity.

Using Profit to Assess Fit

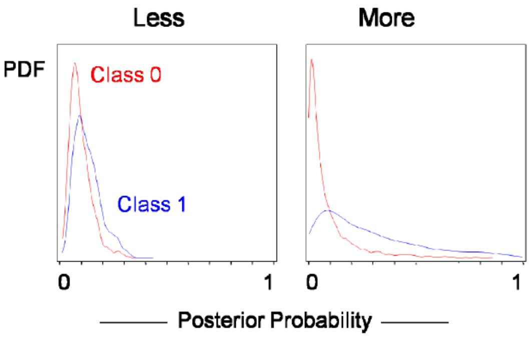


Defining a *profit matrix* can create a useful statistic for measuring classifier performance. The model yields posterior probabilities, and those probabilities classify individuals into likely positives and likely negatives. On the validation data, the behavior of these individuals is known. Hence, it is feasible to calculate each individual's expected profit, and hence it is also feasible to calculate a total profit. This total profit can be used as a model selection and assessment criterion.

If profit information can be used, it permits a more familiar scale for comparing models. Consumers of models might not have a feel for what type of lift they might expect, or what constitutes a good value for sensitivity. Using total or average profit as an assessment statistic might skirt those issues.

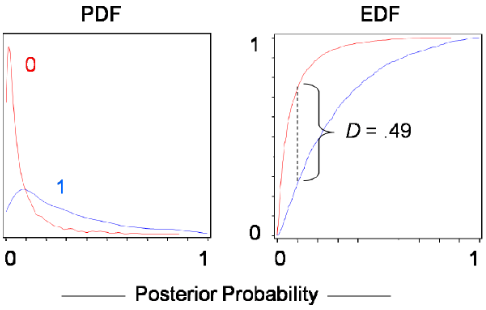
4.4 Overall Predictive Power

Class Separation



Statistics such as sensitivity, positive predictive value, and risk depend on the choice of cutoff value. Statistics that summarize the performance of a classifier across a range of cutoffs can also be useful for assessing global discriminatory power. One approach is to measure the separation between the predicted posterior probabilities for each class. The more that the distributions overlap, the weaker the model is.

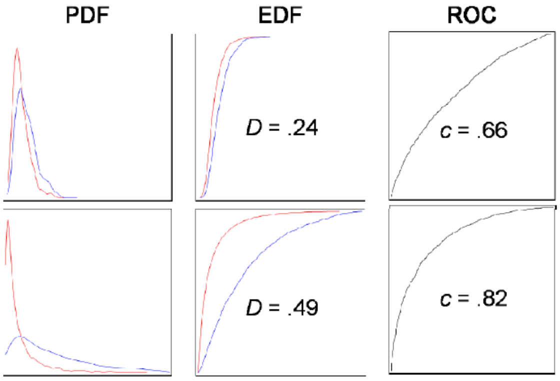
K-S Statistic



The Kolmogorov-Smirnov two-sample test is based on the distance between the empirical distribution functions. The test statistic, D, is the maximum vertical difference between the cumulative distributions. If D equals zero, the distributions are identical. The maximum value of the K-S statistic, 1, occurs when the distributions are perfectly separated. Use of the K-S statistic for comparing predictive models is popular in credit risk modeling.

Area Under the ROC Curve

The K-S statistic is not particularly powerful at detecting location differences because of its generality. The most powerful nonparametric two-sample test is the Wilcoxon-Mann-Whitney test. Remarkably, the Wilcoxon-Mann-Whitney test statistic is also equivalent to the area under the ROC curve.



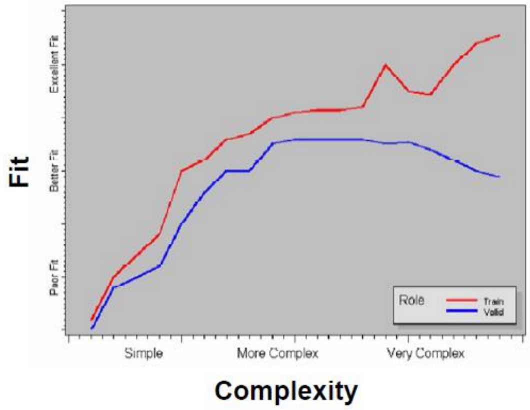
The area under the ROC curve, c, can be determined from the rank-sum in class 1.

A perfect ROC curve would be a horizontal line at one – that is, sensitivity and specificity would both equal one for all cutoffs. In this case, the c statistic would equal one. The c statistic technically ranges from zero to one, but in practice, it should not get much lower than one-half.

Oversampling does not affect the area under the ROC curve because sensitivity and specificity are unaffected. The area under the ROC curve is also equivalent to the Gini coefficient, which is used to summarize the performance of a Lorentz curve.

4.5 Model Selection Plots

Fit versus Complexity



To compare many models, an appropriate fit statistic must be selected. For statistics like average profit, c, and Kolmogorov-Smirnov's D, higher values mean better fitting models. Because the goal of most predictive modeling efforts is a model that generalizes well, these statistics are typically measured on the validation data set.

Typically, model performance follows a fairly straightforward trend. As the complexity increases, the fit on the training data gets better. After a point, the fit might plateau, but on the training data, the fit gets better as model complexity increases. Some of the increase, however, is due to the model identifying vagaries of the training data set. This behavior has been called overfitting. The typical behavior of the validation fit line is an increase followed by a plateau, which might finally result in a decline in performance. The decline is due to overfitting.

ROC and ROCCONTRAST Statements

ROC <*'label'*> <*specification*> <*/ options*>;

ROCCONTRAST <*'label'*> <*contrast*> <*/ options*>;

ROC specifies models to be used in the ROC comparisons. You can specify more than one ROC statement.

ROCCONTRAST compares the different ROC models. You can specify only one ROCCONTRAST statement.