Assessing Classifiers

Xiaogang Su, Ph.D.
Department of Mathematical Sciences
University of Texas at El Paso (UTEP)
xsu@utep.edu

April 5, 2018

Contents

1 Prediction Accuracy **2**
1.1 Optimal Cutoff Point 3

2 ROC Curve **4**
2.1 ROC Curve 4
2.2 Area under ROC Curve (AUC) 5
2.3 Selecting Optimal Cutoff via ROC 8

3 Lift, Percentage of Captured Events **9**

A fitted logistic model would provide predicted probabilities $\hat{\pi}_i = \text{expit}(\mathbf{x}_i^T\widehat{\boldsymbol{\beta}})$'s for $i = 1, \ldots n$. It is preferable to have $\hat{\pi}_i$ computed via cross-validation, ideally, from an independent test data set. We will discuss some common methods for assessing the performance of predicted probabilities. In fact, many other classification tools would provide such predicted probabilities as well. So these assessment methods are applicable to other predictive modeling methods such as decision trees and artificial neural networks.

1 Prediction Accuracy

Let $\hat{\pi}_i$ be the output from a classifier. With logistic regression, these come as the predicted probability for each row in a given data set, which range in (0,1). More generally, the output $\hat{\pi}_i$ does not have to be probabilities; e.g., they could be the linear predictors $\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}$ themselves. Many of techniques discussed below apply to output $\hat{\pi}_i$ in its general form.

To obtain the derived dichotomous variable we must compare $\hat{\pi}_i$ to some cutoff point c. The predicted binary outcome is then given by

$$\hat{y}_i = 1\{\hat{\pi}_i > c\} = \begin{cases} 1 & \text{if } \hat{\pi}_i > c \\ 0 & \text{otherwise} \end{cases}$$

It is not really a big deal whether the rule $\hat{\pi}_i > c$ or $\hat{\pi}_i \geq c$ is used to dichotomize $\hat{\pi}_i$, but this choice would affect some details, e.g., the ends of the ROC curve as we shall see. By default, the cutoff point $c = 0.5$, as suggested by the optimal Bayes classifier, is often used. Another common choice is $c = \bar{y} = \bar{\pi}$, i.e., the proportion of 1's in observed data.

For a fixed threshold c, comparison of the predicted events with the actually observed events leads to a 2×2 table, often termed as the *classification table* or *confusion table*, as shown below.

Table 1: Classification Table by Applying a Cutoff Point

Observed	Predicted Outcome		Row
Outcome	0	1	Total
0	$n_{00}(c)$	$n_{01}(c)$	$n_{0\cdot}$
1	$n_{10}(c)$	$n_{11}(c)$	$n_{1\cdot}$
Column Total	$n_{\cdot 0}(c)$	$n_{\cdot 1}(c)$	n

The dependence of each entry on the threshold c is specified in the table. For example, the total number of cases with predicted $\hat{y} = 0$, $n_{\cdot 0}(c)$, depends on c while the total number of cases with observed $y = 0$, $n_{0\cdot}$, does not. Several useful measures can be defined from the above table.

- *Prediction Accuracy* $= (n_{00} + n_{11})/n$ is the proportion of correctly classified subjects;
- *Sensitivity* $Se = n_{11}/n_{1\cdot}$ is the hit rate or the proportion of correctly classified events out of all events;
- *Specificity* $Sp = n_{00}/n_{0\cdot}$ is the proportion of correctly classified non-events out of all non-events.
- *Positive predictive value* (PPV $= n_{11}/n_{\cdot 1}$) is the proportion of individuals who actually prove to be 1 out of those who has been predicted to be 1.
- *Negative predictive value* (NPV $= n_{00}/n_{\cdot 0}$) is the proportion of individuals who actually prove to be 0 out of those who has been predicted to be 0.

All those measures can be meaningfully interpreted within the setting of medical diagnostic testings where 1 means having the disease and 0 for being disease-free. Both sensitivity and specific are

estimated conditional probabilities. The sensitivity is also called the true positive rate (TPR) while one minus specificity corresponds to the false positive rate (FPR). Besides the $1 - -0$ notion, a common alternative of denoting the binary response value is ± 1. There is a fixed relationship among these three measures:

$$\text{prediction accuracy} = \bar{\pi} \times \text{Se} + (1 - \bar{\pi}) \times \text{Sp},$$

where again $\bar{\pi}$ is the observed proportion of events. By Bayes' rule, PPV and NPV can be expressed in terms of Se and Sp as well:

$$\begin{aligned} PPV &= \hat{\Pr}(y = 1|\hat{y} = 1) = \frac{p_1 Se}{\pi Se + (1 - p_1)(1 - Sp)} \\ NPV &= \hat{\Pr}(y = 0|\hat{y} = 0) = \frac{(1 - p_1)Sp}{(1 - p_1)Sp + p_1(1 - Se)} \end{aligned}$$

where $p_1 = \hat{\Pr}(y = 1)$ denotes the (marginal) proportion of $y = 1$.

For the ICU data, Model I

$$\textbf{Model I}: \quad \text{logit}\left\{\Pr(\texttt{STA} = 1)\right\} = \beta_0 + \beta_1 \cdot \texttt{AGE} + \beta_1 \cdot \texttt{SYS} + \beta_3 \cdot \texttt{LOC}. \tag{1}$$

results in the following classification table using cutoff point .5, with prediction accuracy $(158 + 13)/200 = 85.5\%$, sensitivity $13/40 = 32.5\%$ and specificity $158/160 = 98.75\%$.

Table 2: Classification Table with the ICU Data

Predicted	Observed Death		Row
Death (STA)	0	1	Total
0	158	27	185
1	2	13	15
Column Total	160	40	200

Clearly prediction accuracy alone may not be a very helpful measure for the goodness-of-fit of a logistic model mainly owing to the scale problem: the predicted probabilities from a logistic model are measured on a continuum but the predicted outcomes based on which the prediction accuracy is calculated are binary. Consider two hypothetical models: one predicts all events with probability 0.90 and all nonevents with 0.10, while the other model predicts all events with probability 0.51 and all nonevents with 0.49. Using the default cutoff $c = 0.5$, they would provide the same prediction accuracy. Nevertheless, it is clear that the former model has much better discriminating ability between events and nonevents than the latter one.

1.1 Optimal Cutoff Point

The prediction accuracy depends on the cutoff point employed. Different cutoff points result in different values of misclassification rate, sensitivity, and specificity. Clearly, there is an issue of how to select the optimal cutoff point. Three ways of choosing the cutoff, which really depend on the analytic goals, are in order.

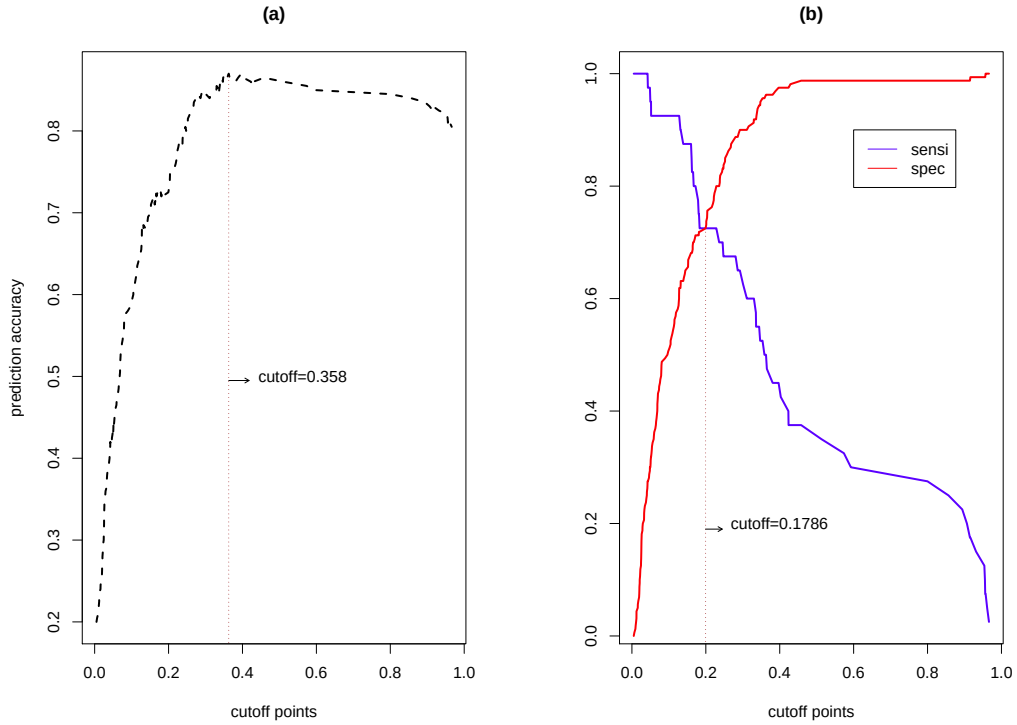

Figure 1: Selection of the Best Cutoff Point with Model I in (1): the ICU Data.

First, one might apply the empirical proportion of events in the original sample. In the ICU example, it would be .20 as 40 out of 200 were deceased. Secondly, if prediction accuracy is the main objective of the analysis, one may select the best threshold as the one that yields the highest prediction accuracy or lowest misclassification rate. Figure 1 (a) plots the prediction accuracy of Model I in (1) versus cutoff for the ICU data. The best choice of the cutoff point is 0.358, at which the prediction accuracy is as high as 87%. The third way is to choose the cutoff point where sensitivity equals specificity. Figure 1(b) plots the both sensitivity and specificity versus cutoff points for the ICU data. The two curves meet at 0.1786.

2 ROC Curve

2.1 ROC Curve

The ROC (receiver operating characteristic) curve originate from signal detection theory that shows how the receiver operates the existence of signal in the presence of noise. It plots the probability of detecting true signal (sensitivity or TPR) and the false signal (1-specificity or FPR) for an entire range of possible cut-points.

Specifically, we sort data in ascending order to $\hat{\pi}_i$:

$$\hat{\pi}_{(1)} < \hat{\pi}_{(2)} < \cdots < \hat{\pi}_{(n')},$$

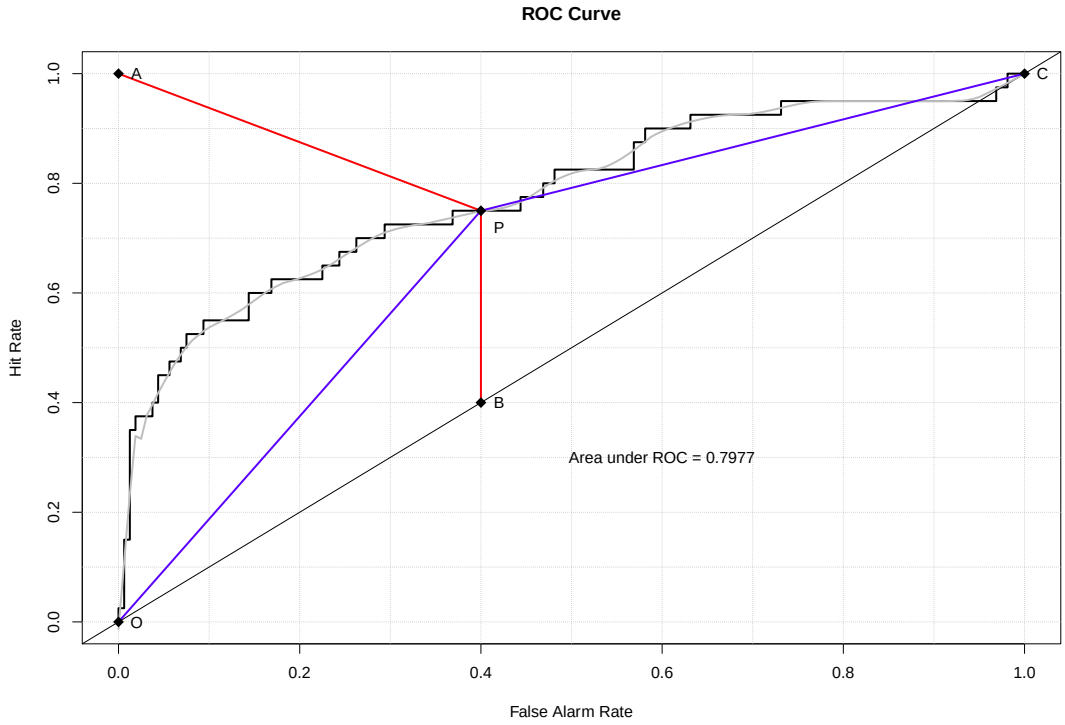

Figure 2: ROC Curve for the ICU Data: Selecting the Best Cutoff $c^\star$.

where n' is the total number of distinct $\hat{\pi}_i$ values. Clearly, the TPR and FPR values only changes at these points of $c = \hat{\pi}_k$ for $k = 1, \ldots, n'$. If $\hat{\pi}_i$ are predicted probabilities from logistic regresison, it is necessary that the minimum $0 < \hat{\pi}_{(1)}$ and the maximum $\hat{\pi}_{(n')} < 1$. But it could occur that $\hat{\pi}_{(1)} = 0$ and $\hat{\pi}_{(n')} = 1$ if they come from other classifier and are scaled to the range [0, 1].

Recall that the dichotomization rule is $\hat{y} = 1$ if $\hat{\pi}_i > c$ with threshold c and 0 otherwise. Starting with $c < \hat{\pi}_{(1)}$, we have $TPR = FPR = 1$. When $c = \hat{\pi}_{(1)}$, TPR and FPR becomes possibly less than 1. When c keeps increasing, both TPR and FPR decrease. When $c = \hat{\pi}_{(n')}$, both TPR and FPR are 0. A typical ROC curve is a step function as plotted in Figure 2; take a closer look at both ends of the step function. Optionally, one may apply linear interpolation for values inbetween (i.e., by connecting these discrete points) or even smooth out the ROC curve.

2.2 Area under ROC Curve (AUC)

The area under the ROC curve (AUC), often termed *c-statistic* or *c-index*, corresponds to the likelihood that an event will have a higher predicted $\Pr(y = 1)$ than a nonevent, which provides a measure of the model's ability to discriminate between those subjects who experience the event versus those who do not. The c-statistic ranges from zero to one. Thumb rules for interpreting AUC (Hosmer and Lemeshow, 2000) are provided below:

- $AUC = .5$: No discrimination (just like flipping a coin);
- $.7 \leq AUC \leq .8$: Acceptable discrimination;

- $.8 \leq AUC \leq .9$: Excellent discrimination;
- $.9 \leq AUC$: Outstanding discrimination;

Figure 2 plots the ROC curve, together with a smoothed version provided by lowess. The area under the ROC curve is 0.7977, showing very good discriminating ability.

AUC can be explicitly computed. AUC has to do the concept of *concordant/discordant pairs*. Suppose we have n observations in the data set. We randomly pick up a pair (i, i') such that $y_i = 1$ and $y_{i'} = 0$. Namely, pairs that have either both 1's on the dependent variable or 0's are not considered. The total number of ways to selecting such a pair is $\binom{n_{0\cdot}}{1} \times \binom{n_{1\cdot}}{1} = n_{0\cdot}\, n_{1\cdot}$, where $n_{0\cdot}$ is the number of negative $(y = 0)$ response and $n_{1\cdot}$ is the number of positive $(y = 1)$ events in the data.

For each pair, we ask the question "Does Case i with the 1-valued response have a higher predicted probability of event (based on model) than Case i' with the 0-valued response?" If the answer is yes, we call this pair *concordant*. If no, the pair is *discordant*. If the two cases have the same predicted value (i.e., $\hat{\pi}_i = \hat{\pi}_{i'}$), we call it a *tie*. Formally, we have the following definition:

Definition 2.1. Given a pair (i, i') that , we call it *concordant* if $(y_i - y_{i'})(\hat{\pi}_i - \hat{\pi}_{i'}) > 0$; discordant if $(y_i - y_{i'})(\hat{\pi}_i - \hat{\pi}_{i'}) < 0$; and *tied* if $y_i \neq y_{i'}$ but $\hat{\pi}_i = \hat{\pi}_{i'}$.

Let n_C denote the total number of concordant pairs; n_D for discordant pairs, and n_T for ties. We have the following proposition for AUC.

Proposition 2.1. *Suppose that we apply dichotomization rule $\hat{y} = 1$ if $\hat{\pi} > c$ for threshold c and plot the ROC curve simply as a step function. The area under the ROC curve (AUC)is given by*

$$AUC = \frac{n_C + n_T}{n_{0\cdot}\, n_{1\cdot}}.$$

Namely, AUC is the proportion of concordant or tied pairs.

Proof. We outline the proof briefly by referring to the graphical illustration in Figure 3. Recall that we have n' distinct sorted predicted probabilities $\hat{\pi}_{(k)}$ for $k = 1, \ldots, n'$. At threshold $c = \hat{\pi}_{(k)}$, we have

$$\begin{aligned} \text{TPR}_k &= \sum_{i=1}^{n} I\{\hat{\pi}_i > \hat{\pi}_{(k)} \,\&\, y_i = +1\}/n_{1\cdot} \\ \text{FPR}_k &= \sum_{i=1}^{n} I\{\hat{\pi}_i > \hat{\pi}_{(k)} \,\&\, y_i = -1\}/n_{0\cdot}. \end{aligned}$$

Note that TPR_k and FPR_k decreases with k to $\text{TPR}_{n'} = \text{FPR}_{n'} = 0$. AUC is essentially the sum of the areas of numerous rectangles. Thus

$$\text{AUC} = \sum_{k=1}^{n'-1} \text{TPR}_k \times (\text{FPR}_k - \text{FPR}_{k+1}) + (1 - \text{FPR}_1), \tag{2}$$

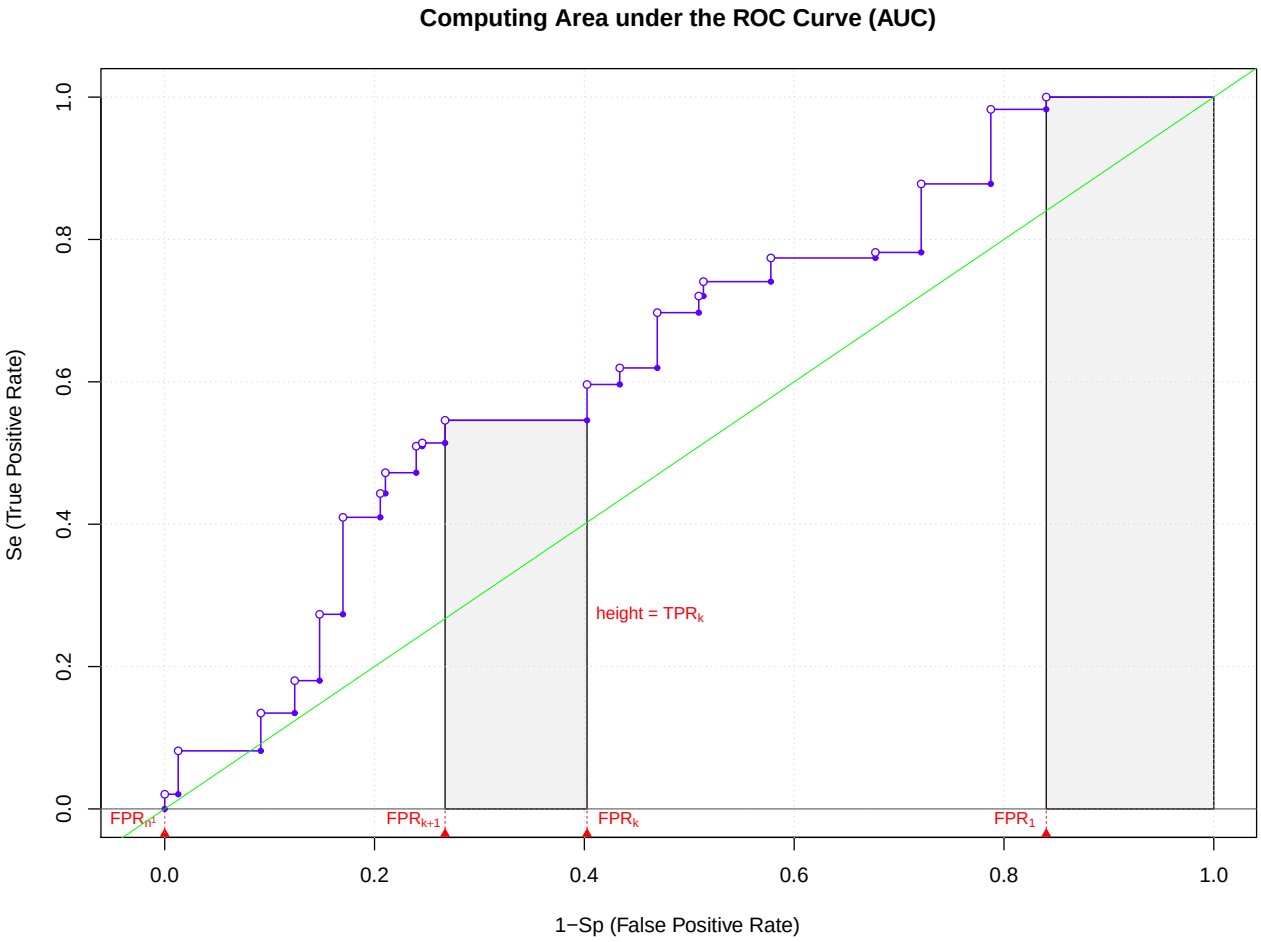

Figure 3: Illustration of AUC Computation.

where the last term corresponds to the rectangle with height 1 on the right of Figure 3 with area

$$\begin{aligned} 1 - \mathrm{FPR}_1 &= 1 - \frac{1}{n_{0\cdot}} \sum_{i=1}^{n} I\left\{\hat{\pi}_i > \hat{\pi}_{(1)} \,\&\, y_i = -1\right\} \\ &= \frac{1}{n_{0\cdot}} \sum_{i=1}^{n} I\left\{\hat{\pi}_i = \hat{\pi}_{(1)} \,\&\, y_i = -1\right\} \end{aligned}$$

since $n_{0\cdot} = \sum_{i=1}^{n} I\left\{\hat{\pi}_i \geq \hat{\pi}_{(1)} \,\&\, y_i = -1\right\}$ and $\hat{\pi}_i \geq \hat{\pi}_{(1)}$ holds for every i. Also, the width of the kth rectangle in the first term is

$$\mathrm{FPR}_k - \mathrm{FPR}_{k+1} = \frac{1}{n_{0\cdot}} \sum_{i=1}^{n} I\left\{\hat{\pi}_i = \hat{\pi}_{(k+1)} \,\&\, y_i = 0\right\}.$$

Plugging the above two equations into (2), after some simple algebra, gives

$$\begin{aligned}
\text{AUC} &= \frac{1}{n_{0\cdot}n_{1\cdot}}\sum_{k=1}^{n'-1}\left[\sum_{i=1}^{n} I\left\{\hat{\pi}_i > \hat{\pi}_{(k)} \& y_i = 1\right\} \cdot \sum_{i'=1}^{n} I\left\{\hat{\pi}_{i'} = \hat{\pi}_{(k+1)} \,\&\, y_{i'} = 0\right\}\right] + \\
&\quad + \frac{1}{n_{0\cdot}}\sum_{i=1}^{n} I\left\{\hat{\pi}_i = \hat{\pi}_{(1)} \,\&\, y_i = -1\right\} \\
&= \frac{1}{n_{0\cdot}n_{1\cdot}}\sum_{k=1}^{n'-1}\left[\sum_{i\neq i'=1}^{n} I\left\{\hat{\pi}_i \geq \hat{\pi}_{(k)} \,\&\, y_i = 1 \,\&\, \hat{\pi}_{i'} = \hat{\pi}_{(k)} \,\&\, y_{i'} = 0\right\}\right] \\
&= \frac{1}{n_{0\cdot}n_{1\cdot}}\sum_{i\neq i'=1}^{n} I\left\{\hat{\pi}_i \geq \hat{\pi}_{i'} \,\&\, y_i = 1 \,\&\, y_{i'} = 0\right\} \\
&= \frac{n_C + n_T}{n_{0\cdot}n_{1\cdot}},
\end{aligned}$$

by using the fact that $I\{\hat{\pi}_{i'} > \hat{\pi}_{(k)}\} = I\{\hat{\pi}_{i'} \geq \hat{\pi}_{(k+1)}\} = I\{\hat{\pi}_{i'} > \hat{\pi}_{(k+1)}\} + I\{\hat{\pi}_{i'} = \hat{\pi}_{(k+1)}\}$. □

It is worth noting that the formula for AUC may vary slightly depending on several variants, e.g., use of linear interpolation and dichotomizing rule. However, in all scenarios, AUC roughly corresponds to the proportion of discordant pairs, measuring if an actually observed event is more likely to be associated with a higher predicted probability when compared to an actually observed nonevent.

2.3 Selecting Optimal Cutoff via ROC

The optimal cutoff point $c^\star$ can also be selected with the aid of ROC curve. Referring to Figure 2, suppose that we pick up one cutoff $c = 0.4$, with corresponding point P on the ROC curve. If we had applied c as the cutoff and hence transferred the predicted $\hat{p}_i$'s into 0 and 1, it would result in a piecewise linear ROC with two segments OP and PC, as highlighted in blue on Figure 2. There are several criteria available.

The first method is to compare with the best possible prediction, as represented by the point A (0,1), in which case the area under ROC curve is 1. The cutoff $c^\star$ minimizes the distance from P to A so that

$$c^\star = \arg\max_c - \sqrt{(1 - \text{Spec}_c)^2 + (1 - \text{Sens}_c)^2}.$$

The second method is to compare with the worst. One measure is represented by the vertical distance PB from P to the reference line $y = x$. It can be shown that

$$|PB| = \text{Sens}_c - (1 - Spec_c).$$

The best threshold $c^\star$ is the one that maximize $|PB|$. This choice of optimal threshold is called the Youden (1950) Index, proposed in cancer biomarker studies. In fact, it can be shown that the area under the curve OPC is

$$AUC = \frac{\text{Sens}_c - (1 - \text{Spec}_c) + 1}{2}.$$

Thus the same choice of $c^\star$ maximizes the area under the curve.

The R package **OptimalCutpoints** (López-Ratón et al., 2014) implements numerous methods for selecting optimal cutpoints, including the one based on Yuden index.

	Non-Cumulative				Cumulative			
Decile	Num of Deaths	Percent of Captured Deaths	Percent of Deaths	Lift	Num of Deaths	Percent of Captured Deaths	Percent of Deaths	Lift
1	12	12/40=30%	12/20=60%	12/4=3	12	12/40=30%	12/20=60%	12/4=3
2	10	10/40=25%	10/20=50%	10/4=2.5	22	22/40=55%	22/40=55%	22/8=2.75
3	2	2/40=5%	2/20=10%	2/4=0.5	24	24/40=60%	24/60=40%	24/12=2
4	2	2/40=5%	2/20=10%	2/4=0.5	26	26/40=65%	26/80=32.5%	26/16=1.63
5	1	1/40=2.5%	1/20=5%	1/4=0.25	27	27/40=67.5%	27/100=27%	27/20=1.35
6	4	4/40=10%	4/20=20%	4/4=1	31	31/40=77.5%	31/120=25.83%	31/24=1.29
7	3	3/40=7.5%	3/20=15%	3/4=0.75	34	34/40=85%	34/140=24.29%	34/28=1.21
8	2	2/40=5%	2/20=10%	2/4=0.5	36	36/40=90%	36/160=22.5%	36/32=1.13
9	1	1/40=2.5%	1/20=5%	1/4=0.25	37	37/40=92.5%	37/180=20.56%	37/36=1.03
10	3	3/40=7.5%	3/20=15%	3/4=0.75	40	40/40=100%	40/200=20%	40/40=1

Figure 4: Worksheet for Computing Percent of Captured Responses, Percent of Responds and Lift Values Based on the Logistic Model I: the ICU Data.

3 Lift, Percentage of Captured Events

There are several other popular measures that describe prediction accuracy from different angles. These include percentage of captured responses, percentage of responses, and lift. We use a worksheet from the ICU data to illustrate the calculations. Again, there are 200 ICU patients in the data set, out of which 40 were dead when released.

We first sort the data according to predicted probabilities $\hat{\pi}_i$'s obtained from fitting the logistic model I in equation (5.15), from the highest to the lowest in a descending order. Then, we divide the data into, say, 10 groups so that each group containing about equally numbered, i.e., 20, observations: patients with the highest 10% of predicted death probabilities go to the 1st group; patients with the second highest 10% of predicted probabilities go to the 2nd group and so on. Within each group, the number of events, i.e., deaths in this example, is recorded. The percentage of captured responses, percentage of responses, and lift values can be computed accordingly. The calculation can be in two modes: cumulatively or non-cumulatively, as demonstrated in the following worksheet. Note that each group contains 20 patients and if the grouping were made in a totally random manner, then each group would have had 4 deaths.

Figure 5 presents the related plots. The noncumulative versions for all these three measures show the same shapes (see the left panels in Figure 5). This is not surprising because, in this example, we happen to have exactly the same number 20 of observations in each group. The plot of lift values, often termed the lift chart, shows how much better the model can do when compared to the null model that corresponds to lift value of 1. The plot of percentages of the captured responses is particularly useful in business decisions. For example, when planning a mailing campaign, let the binary response indicate whether or not a customer would respond to the solicitation mail. The company really does not have to mail out all the solicitation packages, given that some people would be very unlikely to respond and both the mail and the mailing have cost. In order to capture a majority of potential responders, they just need to focus on the top deciles in the plot of the captured responses.

As a final note, calculation of all the aforementioned measures is preferably based on an in-

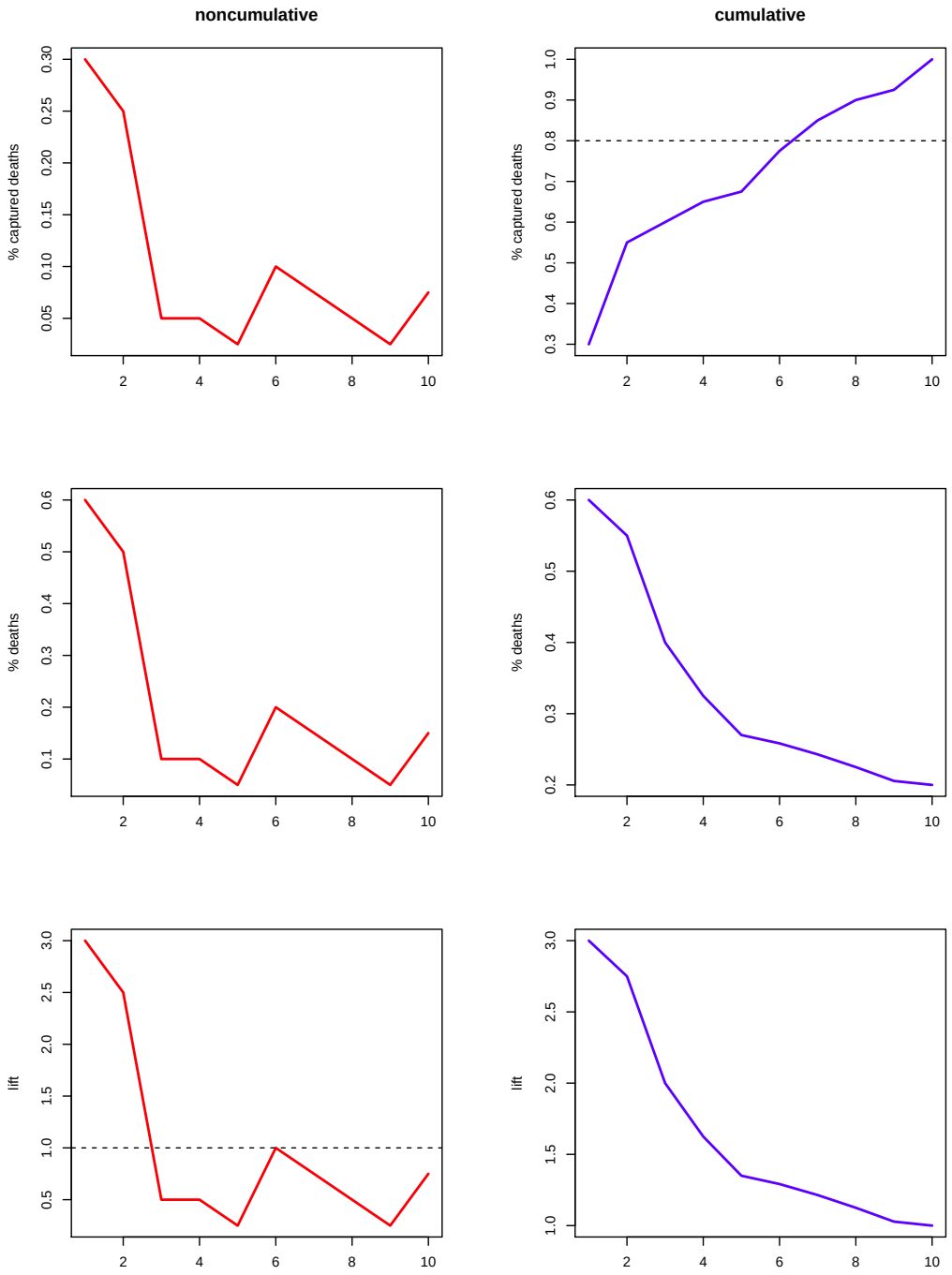

Figure 5: Plot of Percent of Captured Responses, Percent of Responds and Lift Based on Model I in (1): the ICU Data. The red lines correspond to noncumulative values while the blue lines are for cumulative values.

dependent data set in order to have an honest evaluation on the model performance. Otherwise, the results tend to be overoptimistic to some extent.In the cases when an independent data set is not available due to limited sample size, one can resort to resampling techniques such as v-fold cross-validation or bootstrapping.

References

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *Elements of Statistical Learning*, 2nd Edition. Chapman and Hall.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logisitc Regression.* New York: John Wiley & Sons, Inc.

López-Ratón, M., Rodriguez-Álvarez, M. X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). **OptimalCutpoints**: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software*, **61**, Issue 8. URL https://www.jstatsoft.org/article/view/v061i08/v61i08.pdf

Su, X. G. and Gau, G. (2010). Predictive Modeling in Healthcare, Chapters 8-12 for *Risk Adjustment and Predictive Modeling*, Ed. by Duncan, I.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, **3**: 32–35.