Predictive Modeling

Department of Mathematical Sciences
Qiuying Sha, Professor

Michigan Tech

Part III. Classification Models

Chapter 11. Measuring Performance in Classification Models

- In the previous part of this book we focused on building and evaluating models for a continuous response.
- We now turn our focus to building and evaluating models for a categorical response.
- Although many of the regression modeling techniques can also be used for classification, the way we evaluate model performance is necessarily very different since metrics like RMSE and R^2 are not appropriate in the context of classification.
- We discuss **metrics** for evaluating classification model performance using **statistics and visualizations**.

11.2 Evaluating Predicted Classes

Confusion matrix: a common method for describing the performance of a classification model.

Table 11.1: The confusion matrix for the two-class problem ("events" and "nonevents." The table cells indicate number of the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)

Predicted	Observed	
	Event	Nonevent
Event	TP	FP
Nonevent	FN	TN

Table 11.1 shows an example when the outcome has two classes.

- **Diagonal cells** denote cases where the classes are correctly predicted.
- The **off-diagonals** illustrate the number of errors for each possible case.

The simplest metric is the overall **accuracy rate** (or, the **error rate**).

- That has the most straightforward interpretation.

There are a few disadvantages to using this statistic.

- First, overall accuracy counts make **no distinction about the type of errors being made**.

  In spam filtering, the cost of erroneous deleting an important email is likely to be higher than incorrectly allowing a spam email past a filter.

- Second, one must consider the natural frequencies of each class.
  - For example, in the USA, pregnant women routinely have blood drawn for alphafeto protein testing, which attempts to detect genetic problems such as Down Syndrome.
  - Suppose the rate of this disorder in fetuses is approximately 1 in 800 or about one-tenth of one percent.
  - A predictive model can achieve almost perfect accuracy by predicting all samples to be negative for Down Syndrome.

Rather than calculate the overall accuracy, other metrics can be used that take into account the class distributions of the training set samples.

The *Kappa statistic* (also known as Cohen's Kappa) (Cohen 1960) takes into account the accuracy that would be generated simply by chance. The form of the statistic is

$$Kappa = \frac{O - E}{1 - E}$$

where *O* is the observed accuracy; *E* is the expected accuracy based on the marginal totals of the confusion matrix.

- A value of 0 means there is no agreement between the observed and predicted classes other than what would expected by chance.
- A value of 1 indicates perfect concordance of the model prediction and the observed classes.
- Negative values indicate that the prediction is in the *opposite* direction of the truth, but large negative values seldom occur.

Kappa values within 0.30 to 0.50 indicate reasonable agreement.

The Kappa statistic can also be extended to evaluate concordance in problems with more than two classes.

See supplement: Kappa from http://en.wikipedia.org/wiki/Cohen's_kappa

Two-Class Problems

- For two classes, there are additional statistics that may be relevant when one class is interpreted as the event of interest (such as Down syndrome in the previous example).

- The *sensitivity* of the model is the rate that the event of interest is predicted correctly for all samples having the event, or

$$Sensitivity = \frac{\#\text{ samples with the event } and \text{ predicted to have the event}}{\#\text{ samples having the event}}$$

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

  - The sensitivity is sometimes considered the *true positive rate* since it measures the accuracy in the event population.

The *specificity* is defined as the rate that nonevent samples are predicted as nonevents, or

$$Specificity = \frac{\#\text{ samples without the event } and \text{ predicted as nonevents}}{\#\text{ samples without the event}}$$

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

- The false-positive rate = 1-specificity.

- Assuming a fixed level of accuracy for the model, there is typically a trade-off to be made between the sensitivity and specificity.

  - Intuitively, increasing the sensitivity of a model is likely to incur a loss of specificity, since more samples are being predicted as events.

In Chap. 4 we introduced the credit scoring example.

- Since the performance of the two models were roughly equivalent, the logistic regression model was favored due to its simplicity.

- Using the previously chosen test set of 200 customers, Table 11.2 shows the confusion matrix associated with the logistic regression model.

Table 11.2: Test set confusion matrix for the logistic regression model training with the credit scoring data from Sect. 4.5

Predicted	Observed	
	Bad	Good
Bad	24	10
Good	36	130

- The overall accuracy was 77%.
- The test set had a Kappa value of 0.375, which suggests moderate agreement.
- If we choose the event of interest to be a customer with bad credit, the sensitivity from this model would be estimated to be 40% and the specificity to be 92.9%.

- The model has trouble predicting when customers have bad credit.
  - This is likely due to the imbalance of the classes and a lack of a strong predictor for bad credit.

- The most common method for combining sensitivity and specificity into a single value uses the receiver operating characteristic (ROC) curve, discussed below.

11.3 Evaluating Class Probabilities

One approach to using the probabilities to compare models is ROC curves.

Receiver Operating Characteristic (ROC) Curves

- ROC curves (Altman and Bland 1994; Brown and Davis 2006; Fawcett 2006) were designed as **a general method** that, given a collection of continuous data points, determine an **effective threshold** such that values **above the threshold are indicative of a specific event**.

- This tool will be examined in Chap. 19, but here, we describe **how the ROC curve can be used for determining alternate cutoffs for class probabilities**.

- For the credit model test set previously discussed,
  - The sensitivity was poor for the logistic regression model (40%).
  - The specificity was fairly high (92.9%).
  - These values were calculated from classes that were determined with the **default 50% probability threshold**.
  - Can we improve the sensitivity by **lowering the threshold** to capture more true positives?
  - Lowering the threshold for **classifying bad credit to 30%** results in a model with improved **sensitivity (60%)** but decrease specificity **(79.3%).**
  - In Fig. 11.3, we see that **decreasing the threshold** begins to **capture more of the customers with bad credit** but also begins to encroach on the bulk of the customers with good.

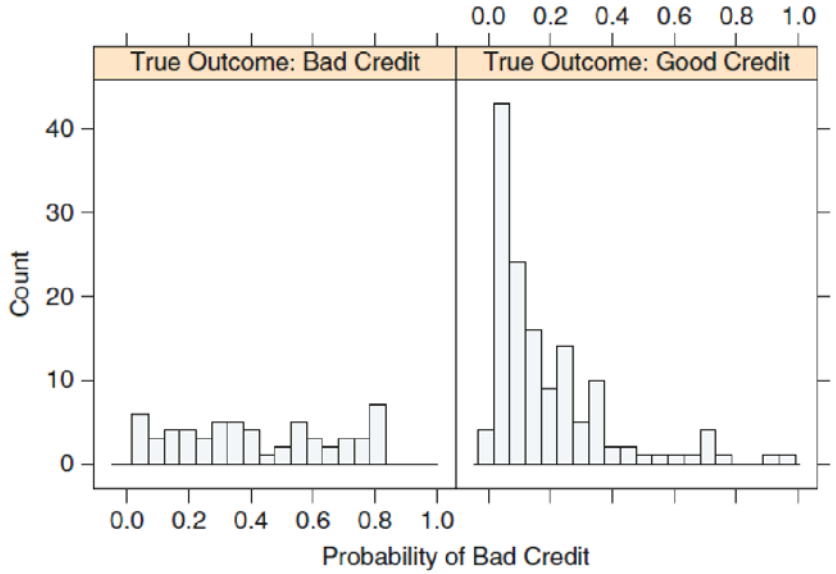


Fig. 11.3: Histograms for a set of probabilities associated with bad credit. The two panels split the customers by their true class.

- The ROC curve is created by **evaluating the class probabilities** for the model **across a continuum of thresholds.**

- For **each candidate threshold**, the resulting **true-positive rate** (i.e., the sensitivity) and the **false-positive rate** (one minus the specificity) are plotted against each other.

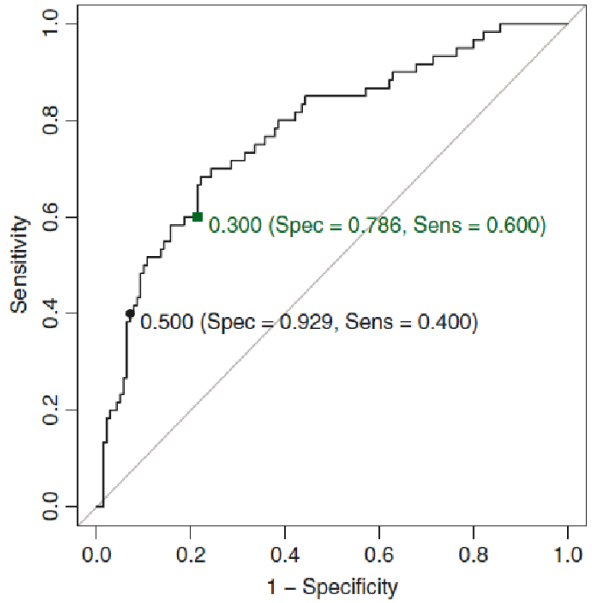


Fig. 11.6: A receiver operator characteristic (ROC) curve for the logistic regression model results for the credit model. The dot indicates the value corresponding to a cutoff of 50% while the green square corresponds to a cutoff of 30% (i.e., probabilities greater than 0.30 are called events)

Figure 11.6 shows the results of this process for the credit data.

The between (0, 0) and the 50% threshold is steep, indicating that the sensitivity is increasing at a greater rate than the decrease in specificity.

When the sensitivity is greater than 70%, there is a more significant decrease in specificity than the gain in sensitivity.

This plot is a helpful tool for **choosing a threshold** that appropriately **maximizes the trade-off between sensitivity and specificity**.

- The **ROC curve** can also be used for a **quantitative assessment of the model**.
  - A perfect model that completely separates the two classes would have 100% sensitivity and specificity.
  - Graphically, the ROC curve would be a single step between (0, 0) and (0, 1) and remain constant from (0, 1) to (1, 1). The area under the ROC curve for such a model would be one.
  - A **completely ineffective model** would result in an ROC curve that closely follows the **45° diagonal line** and would have **an area under the ROC curve of approximately 0.50**.

- The **optimal model** should be shifted towards the **upper left corner** of the plot.

Alternatively, the model with **the largest area under the ROC curve** would be the most effective.

For the credit data, the logistic model had an estimated **area under the ROC curve of 0.78** with a 95% confidence interval of (0.7, 0.85) determined using the bootstrap confidence interval method.

- A **disadvantage of using the area under the curve (AUC)** to evaluate models is that it **obscures information**.

  - For example, when comparing models, it is common that no individual ROC curve is uniformly better than another (i.e., the curves cross).

  - By summarizing these curves, there is **a loss of information**, especially if one particular area of the curve is of interest.

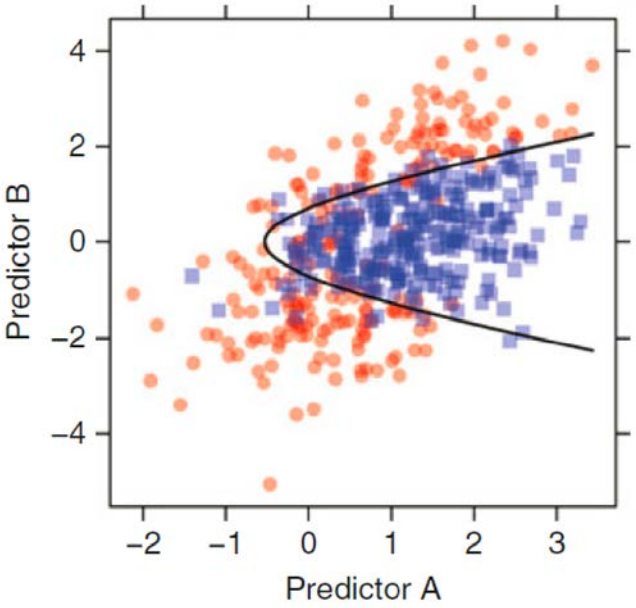
For example, one model may produce a steep ROC curve slope on the left but have a lower AUC than another model. If the lower end of the ROC curve was of primary interest, then AUC would not identify the best model.

- The **partial area under the ROC curve** (McClish 1989) is an alternative that on specific parts of the curve.

- The ROC curve is **only defined for two-class problems** but has been extended to **handle three or more classes**.

  - Hand and Till (2001), Lachiche and Flach (2003), and Li and Fine (2008) use different approaches extending the definition of the ROC curve with more than two classes.

Data simulation:



For two classes (classes 1 and 2) and two predictors (*A* and *B*), the true probability (*p*) of the event is generated from the equation:

$$\log\left(\frac{p}{1-p}\right) = -1 - 2A - .2A^2 + 2B^2$$

Figure 11.1 shows a simulated test set along with the a contour line for a $p = 0.50$ event probability.

Two models were fit to the training set: quadratic discriminant analysis (QDA, Sect. 13.1) and a random forest model (Sect. 14.4).