ROC Curve and Lift Curve : GRAPHS F0R   GOODNESS   OF   FIT

Reference
1. StatSoft, Inc. (2011). STATISTICA (data analysis software system), version 10. www.statsoft.com.
2. JMP, Version 9. SAS Institute Inc., Cary, NC, 1989-2009.
3. A. Moise, R. Salamon, D. Commenges, B. Clément (1986),
   Revue Épidémiologie et Santé Publique, vol. 34, pp 209-217

For continuous responses: the Actual by Predicted plot how well the model fits the models. For a perfect fit, all the points would be on the diagonal.

ROC Curve

For categorical responses the ROC curve is used. The classical definition of ROC curve involves the count of True Positives by False Positives as you accumulate the frequencies across a rank ordering. The True Positive y-axis is labeled "Sensitivity" and the False Positive X-axis is labeled "1-Specificity". The idea is that if you slide across the rank ordered predictor and classify everything to the left as positive and to the right as negative, this traces the trade-off across the predictor's values.

To generalize for polytomous cases (more than 2 response levels), **Partition creates an ROC curve for each response level versus the other levels**. If there are only two levels, one is the diagonal reflection of the other, representing the different curves based on which is regarded as the "positive" response level.
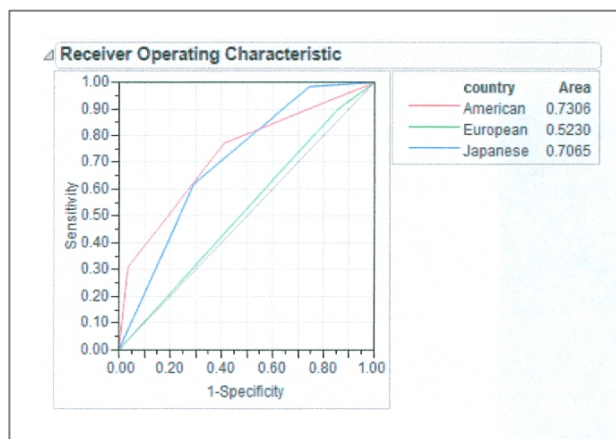

ROC curves are nothing more than a curve of the sorting efficiency of the model. The model rank-orders the fitted probabilities for a given Y-value, then starting at the lower left corner, draws the curve up when the row comes from that category, and to the right when the Y is another category. In the following picture, the Y axis shows the number of Y's where Y=1 and the X axis shows the number of Y's where Y=0.
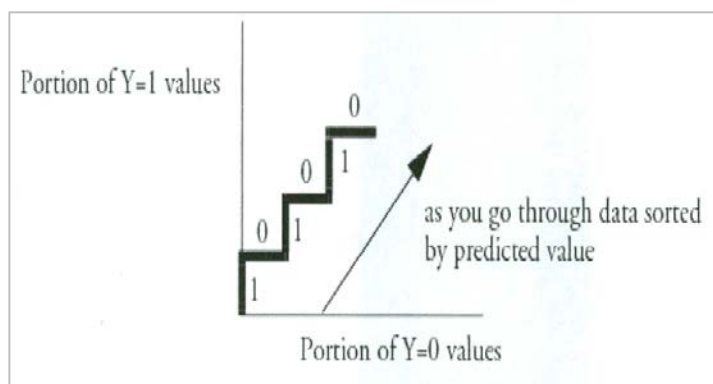
If the model perfectly rank-orders the response values, then the sorted data has all the targeted values first, followed by all the other values. The curve moves all the way to the top before it moves at all to the right.

ROC for Perfect Fit
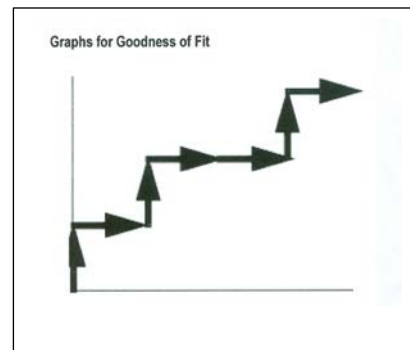
If the model does not predict well, it wanders more or less diagonally from the bottom left to top right. In practice, the curve lifts off the diagonal. The area under the curve is the indicator of the goodness of fit, with 1 being a perfect fit.

If a partition contains a section that is all or almost all one response level, then the curve lifts almost vertically at the left for a while. This means that a sample is almost completely sensitive to detecting that level. If a partition contains none or almost none of a response level, the curve at the top will cross almost horizontally for a while. This means that there is a sample that is almost completely specific to not having that response level. Because partitions contain clumps of rows with the same (i.e. tied) predicted rates, the curve actually goes slanted, rather than purely up or down.

For polytomous cases, you get to see which response categories lift off the diagonal the most. In the CarPoll example above, the European cars are being identified much less than the other two categories. The American's start out with the most sensitive response (Size(Large)) and the Japanese with the most negative specific (Size(Large)'s small share for Japanese).

A ROC curve can be used to evaluate the goodness of fit for a binary classifier. It is a plot of the true positive rate (rate of events that are correctly predicted as events) against the false positive rate (rate of nonevents predicted to be events) for the different possible cutpoints. A ROC curve demonstrates the following:
- The trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

ROC. Use the options in this group box to create **ROC (Receiver Operating Characteristic) curves** and spreadsheets for the active neural networks and specified samples. Note that this group box is available only for binary classification problems.

ROC curve. Click this button to create a line plot of the ROC curve for any number of active neural networks in one window given the data samples. Same information can be generated in spreadsheet format using the option below.

ROC spreadsheet. Click this button to create two spreadsheets, one containing data (i.e., 1-specifivity versus sensitivity) and the other containing estimates of the ROC areas and thresholds for each active neural network and specified data samples.

Lift Curves

A lift curve shows the same information as an ROC curve, but in a way to dramatize the richness of the ordering at the beginning. The Y-axis shows the ratio of how rich that portion of the population is in the chosen response level compared to the rate of that response level as a whole. For example, if the top-rated 10% of fitted probabilities have a 25% richness of the chosen response compared with 5% richness over the whole population, the lift curve would go through the X-coordinate of 0.10 at a Y-coordinate of 25% / 5%, or 5. All lift curves reach (1, 1) at the right, as the population as a whole has the general response rate.

Sorted from highest predicted rate to lowest predicted rate
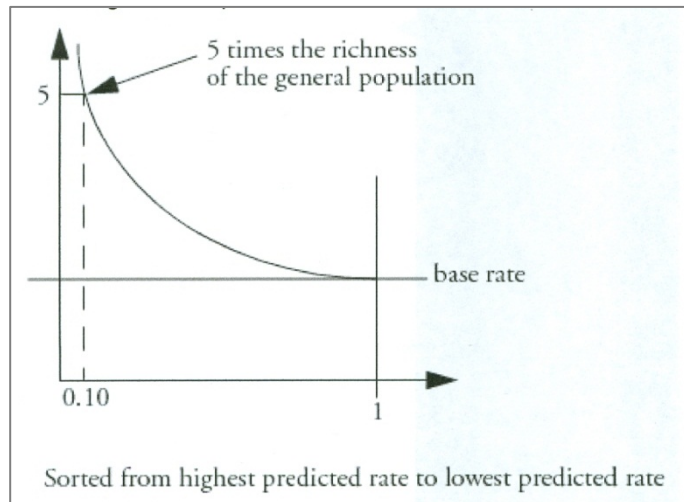
In problem situations where the response rate for a category is very low anyway (for example, a direct mail response rate), the lift curve explains things with more detail than the ROC curve.
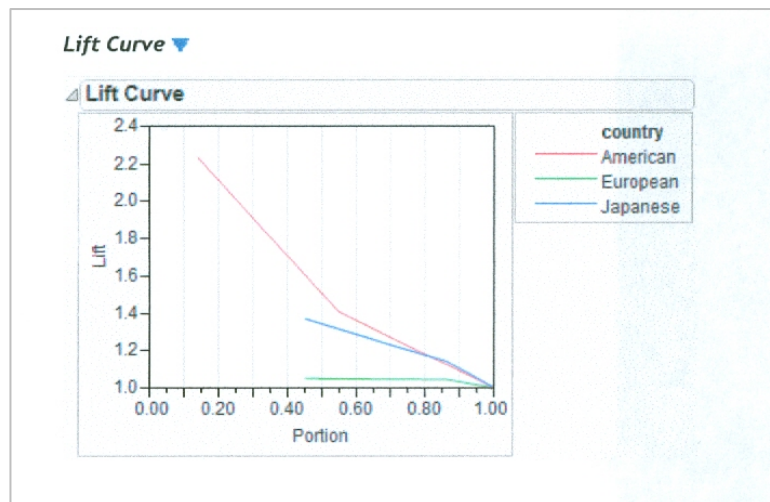
Liftcharts Tab (SANN of Statistica)

Select the *Liftcharts* tab of the *SANN - Results* dialog to access the options described here. This tab is only available for classification models. For information on the options that are common to all tabs, see *SANN - Results*. Note that for all lift chart types, you can include cases in the *Train*, *Test*, and/or *Validation* subsets by selecting the appropriate check boxes in the *Sample* group box. For example, to view a gains chart of all categories for only the *Validation* subset, select the *Gains chart* option button, select *All* in the *Category* list, select the *Validation* check box in the *Sample* group, and click the *Lift chart* button. Only the results for cases in the validation sample will be plotted.

Type. The options in this group box are used to create lift charts and gains charts for the categories of the target variables and for the current model. **Use these charts to evaluate and compare the utility of the model for predicting the different categories or classes for the categorical target variable.** Select the option button that specifies the type of chart and the scaling for the chart you want to compute.

Gains chart. Select this option button to compute a gains chart. This chart shows the percentage of observations correctly classified into the chosen category (see *Category of response* below) when taking the top *x* percent of cases from the sorted (by classification probabilities) data file.

For example, this chart can show you that by taking the top 20 percent (shown on the x-axis) of cases classified into the respective category with the greatest certainty (maximum classification probability), you would correctly classify almost 80 percent of all cases (as shown on the vertical y-axis of the plot) belonging to that category in the population. In this plot, the baseline random classification (selection of cases) would yield a straight line (from the lower-left to the upper-right corner), which can serve as a comparison to gauge the utility of the respective models for classification.

Lift chart (response %). Select this option button to compute a lift chart where the vertical y-axis is scaled in terms of the percentage of all cases belonging to the respective category. As in the gains chart, the x-axis denotes the respective top *x* percent of cases from the sorted (by classification probabilities) data file.

Lift chart (lift value). Select this option button to compute a lift chart where the vertical y-axis is scaled in terms of the lift value, expressed as the multiple of the baseline random selection model.

For example, this chart can show you that by **taking the top 20 percent** (shown on the x-axis) of cases classified into the respective category with the greatest certainty (maximum classification probability), you would end up with a sample that has almost 4 times as many cases belonging to the respective category when compared to the baseline random selection (classification) model.

Category. Select the response category for which to compute the gains and/or lift charts. You can chose to produce lift charts for a single or all categories.

Cumulative. Select this check box to show in the chosen lift and gains charts the cumulative percentages, lift values, etc. Clear this check box to show the simple (noncumulative) values.

Lift chart. Click this button to create the chart as specified via the options above.