Model Accuracy Measures

Lecture 7

05/08/2021

Judging the fit of a Logistic Regression

- The overall fit of a multiple linear regression model is judged by the value of R^2 from the fitted model.
- For logistic regression, some ad hoc measures have been proposed which are based on the ratio of likelihoods.
- One method to judge how well the model is doing we will determine the number of observations in the sample that the model is classifying correctly.
- The primary objective of logistic regression is to solve classification problems based on the predicted probability values.
- Once you fit the logistic model to the data, calculate the fitted logits and from the fitted logits, we will calculate the fitted probabilities for each observation.

Contd…

- To classify the observations, the analyst has to first decide the classification cut-off probability P_c . Whenever the predicted probability of an observation is less than the cut-off probability, then the observation is classified as negative (Yi=0) and if the predicted probability is greater than or equal to cut-off probability, then the observation is classified as positive (Yi=1). That is

$$Y_i = \begin{cases} 0 & if\ P(Y_i = 1) < P_c \\ 1 & if\ P(Y_i = 1) \geq P_c \end{cases}$$

Classification Table

- The classification table in a logistic regression model output is a table that provides accuracy of the logistic regression model (accuracy of classifying positives and negatives) for a chosen classification cut-off probability.
- Different cut off values have been suggested in the literature.
- Commonly used methods for selecting the cut-off probability are

1. Youden’s index
2. Cost-based approach

- By default , we consider if the fitted probability of an observation is greater than or equal to 0.5, we will assign it to Group 1(Y=1) and if less than 0.5, we will classify it in Group 0(Y=0).
- We will then determine what proportion of the data is classified correctly. A high proportion of correct classification indicate that the logistic model is working well. A low proportion of correct classification indicates poor performance.

Classification Table(Confusion matrix)

- The above discussed process will be applied to the original data that was used to fit the logistic regression model.
- The resultant table takes the form given below.

		Predicted Result		
		Success	Failure	
Actual	Success	n_{SS}	n_{SF}	$n_{S.}$
Result	Failure	n_{FS}	n_{FF}	$n_{F.}$
		$n_{.S}$	$n_{.F}$	n

- The proportion of observations correctly classified is $(n_{SS} + n_{FF})/n$.

Sensitivity and Specificity (Notations)

- In classification, the model performance is often measured using concepts such as sensitivity, specificity, precision and F-score.
- The ability of the model to correctly classify positives and negatives is called sensitivity(also known as recall or true positive rate)and specificity(also known as true negative rate).
- Let TN,FN,FP and TP represent the numbers of true negatives, false negatives, false positives and true positives resp.

Also let

- TAN : Total Actually Negative= TN + FP
- TAP : Total Actually Positive= FN + TP

- Sensitivity = Model classifies Yi as positive/Yi is positive

  =Number of true positives/Total actually positive.

  = TP/(FN+TP)=TP/TAP

- Specificity = Model classifies Yi as negative/Yi is negative

  = Number of true negatives/ Total actually negative.

  =TN/(TN+FP)=TN/TAN

Precision and F-score

- Precision is the conditional probability that the actual value is positive given that the prediction by the model is positive. Mathematically precision is given by

$$Precision = \frac{TP}{TP + FP}$$

- F-score is a measure that combines the precision and recall (Harmonic mean between precision and recall). Mathematically F-score is given by

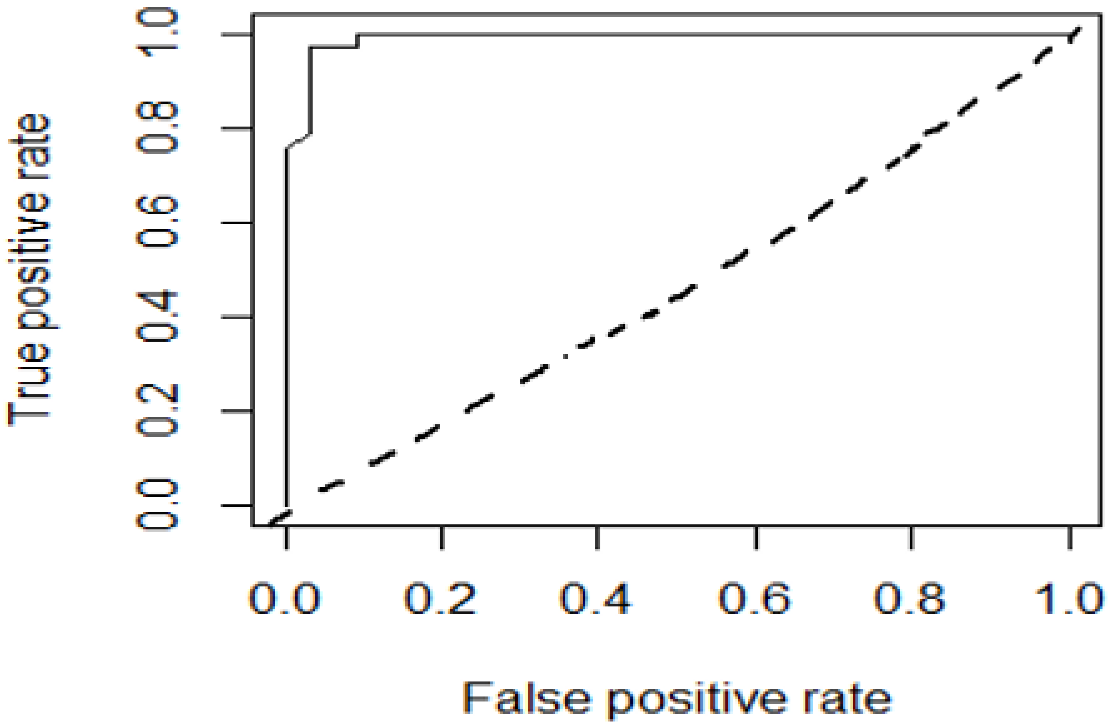
$$F - score = \frac{2 * Recall * Precision}{Recall + Precision}$$

Receiver Operating Characteristic (ROC)

- ROC can be used to understand the overall performance of a logistic regression model(and in general, of classification models) and used for model selection.
- Given a random pair of positive and negative class records, ROC gives the proportions of such pairs that will be correctly classified.
- In a Receiver Operating Characteristic (**ROC**) **curve,** the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points.
- That is ROC is a plot between sensitivity(true positive rate) on the vertical axis and 1-specificity(false positive rate) on the horizontal axis.
- For a perfect classification model, we would have both sensitivity and specificity =1.

ROC

- The diagonal line in the figure represents the case of not using a model(no discrimination between positive and negative)
- Sensitivity and /or Specificity are likely to change when the cut-off probability is changed.
- The line above the diagonal line captures how sensitivity and (1-specificity) change when the cut-off probability is changed.
- The closer the curve follows the left-hand 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.



Area Under ROC Curve (AUC)

- AUC is the proportion of concordant pairs (a pair of positive and negative observations for which the model has a cut-off probability to classify both of them correctly) in the data if the model is used for classification.
- AUC is one of the criteria used for final model selection; higher AUC is assumed to be better model.
- As a rule of thumb, AUC of atleast 0.7 is required for practical application of the model. Caution should be exercised while selecting models based on AUC, especially when the data is imbalanced.
- In case of imbalanced dataset, AUC may be very high(greater than 0.9); however either sensitivity or specificity values may be poor.

Gain chart and Lift chart

- Gains charts and Lift charts are graphical evaluative methods for assessing and comparing the usefulness of classification models.
- These are used in business contexts such as target marketing.
- In target marketing, customer’s response to campaign are usually very low. The organization incurs cost for each customer contact and hence would like to minimize the cost of marketing campaign; and at the same time achieve the desired response level from the customers.
- Suppose that a financial leading firm is interested in identifying the high income persons to put together a targeted marketing campaign for a new platinum credit card.
- It is much better to apply demographic information that the company may have about the list of contacts , build a model to predict which contacts will have high income, and restrict the canvassing to these contacts classified as high income.
- The cost of marketing program will then be much reduced and the response rate may be higher.

The gain chart and lift charts are obtained using the following steps.

1. Predict the probability Y=1(positive) using the logistic regression model (LR)and arrange the observation in the decreasing order of predicted probability.[i.e., P(Y=1)].
2. Divide the datasets into deciles . Calculate the number of positives (Y=1) in each decile and cumulative number of positives up to a decile.
3. Gain is the ratio between cumulative number of positive observations up to a decile to total number of positive observations in the data. Gain chart is a chart between gain on the vertical axis and the decile on the horizontal axis.

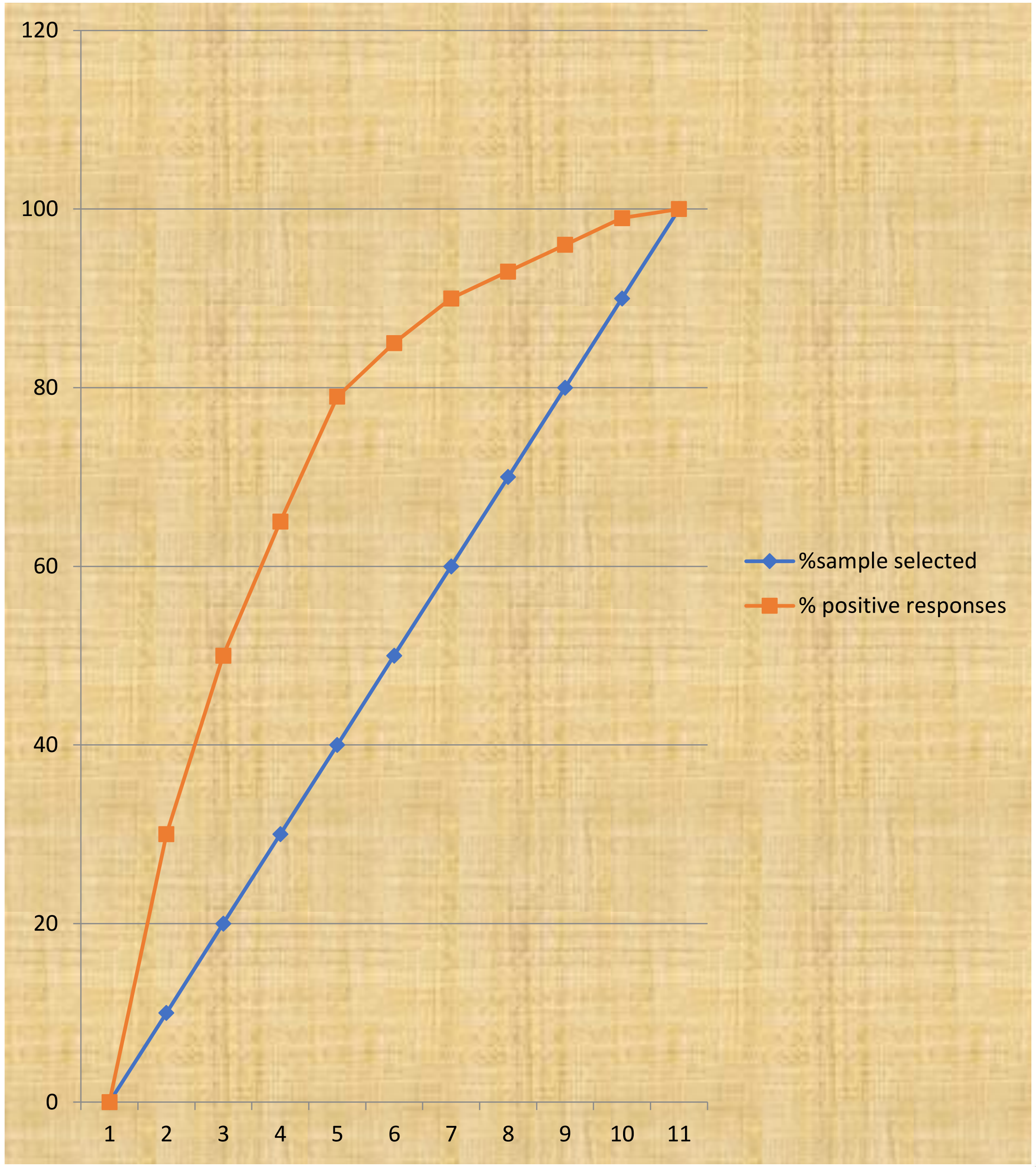
$$\text{Gain}= \frac{\textit{Cumulative number of positive observations upto decile } i}{\textit{Total number of positive observations in the data}}$$

4. Lift is the ratio of number of positive observations up to decile i using the LR model to the expected number of positives up to that decile i based on a random model (not using a model). Lift chart is the chart between lift on vertical axis and the corresponding decile on the horizontal axis.

$$\text{Lift}= \frac{\textit{Cumulative number of positive observations upto decile } i \textit{ using LR model}}{\textit{Cumulative number of positive observations up to decile } i \textit{ based on a random model}}$$

Cumulative Gains Chart

- The baseline trend shows the response that will be obtained as a result of not using the model. It is linear and shows that targeting x% of the population will result in identifying x% of the responders/purchasers.
- The curve shows the improvement in identifying the responders/purchasers as a result of applying the propensity model. In the chart, the model enables identifying 85% of the responders /purchasers by targeting only 50 % of the population compared to identifying only 50% of the responders/purchasers without using the model.
- Propensity model will enable realizing substantial savings in terms of being able to target lesser number of entities.


%sample selected
% positive responses