Performance and Economic Evaluation of Fraud Detection Systems

GCX Advanced Analytics LLC

Fraud risk managers are interested in detecting and preventing fraud, but when it comes to making a business case to management for a new, analytics-based, detection system, it is frequently difficult to quantify the benefits of the system. Even if all the other parts of the system are fine from a software application perspective (alert management, case management, list management, data acquisition, reporting, and so on), if the core analytics are not performing in an economically beneficial way, the rest of the system is not helpful in reducing fraud losses.

In this paper, GCX shows how to evaluate the detection performance of the analytics component or model, and then translate that performance into an economic benefit (or cost). Further, we show how to optimize the operational parameters of the detection system such that the net economic benefit of the system is maximized.

Fraud detection systems are *binary classifiers*; i.e. given an example belonging to one of two classes ("*fraud*" or "*not fraud*") the classifier assigns the example to one category, either correctly or incorrectly. A perfect classifier identifies all the frauds, and does not generate any false positives. Practical classifiers have error rates associated with this classification process.

Receiver Operating Characteristic Curves[1]. The graph of the relationship between detection rate and false positive rate for a system is referred to as the receiver operating characteristic (ROC) curve, an example of which is shown in Figure 1 below.

Economic measures of effectiveness, ultimately, where we factor in the cost of reviewing alerts, the cost of a wrong decision, the cost of undetected fraud, and the fixed and variable costs of operating the system become the important measures.

Receiver Operating Characteristic Curve Analysis

An excellent method of performance evaluation of detector or binary classifier systems leverages the receiver operating characteristic (ROC) curve, often to compare one classifier against another. The ROC is a plot in [0,1],[0,1] of the true positive rate (detection rate) on the y-axis versus the false positive rate on the x-axis. An example illustrating the relationship between the contingency table, the population distributions, and the ROC is shown in Figure 1.

The ROC generates a curve that for any reasonable classifier (i.e. better than random; a random classifier will generate a diagonal line in ROC space) will capture an area of at least 0.5. This is the *area under curve* (AUC) metric of an ROC curve. *All else equal, a classifier with a greater AUC is generally preferred to another one with a lower AUC.*

Best and Worst Case ROCs. The two extreme cases of ROCs are for the *perfect classifier* and the *random classifier*. The perfect classifier would detect all of the frauds and generate no false positivies, so its ROC curve would go straight up from [0,0] to [1,0] and then over to [1,1]. The AUC of the perfect classifier is 1.0 then. A random classifier is one obtained by assigning example to fraud or not-fraud based on the outcome of a fair coin flip. This procedure generates equal detection and false positive rates regardless of any other parameter. The ROC of the random classifier is a diagonal line going from [0,0] to [1,1] and has

[1] ROC methods were first developed by the US Army to characterize the operation of radar systems with respect to correctly differentiating friendly from foe aircraft. This is a signal theory application, and has been applied in the medical diagnostic testing area successfully for the past 20 years. ROC methods are now being applied in the area of intrusion detection and fraud detection systems.

an AUC of 0.5. If a classifier somehow has an AUC less than 0.5, it can be inverted to obtain a better classifier with an AUC of 1 minus the original AUC[2].

An example of a ROC curve from a GCX credit card fraud detection models appears below in Figure 1 below.

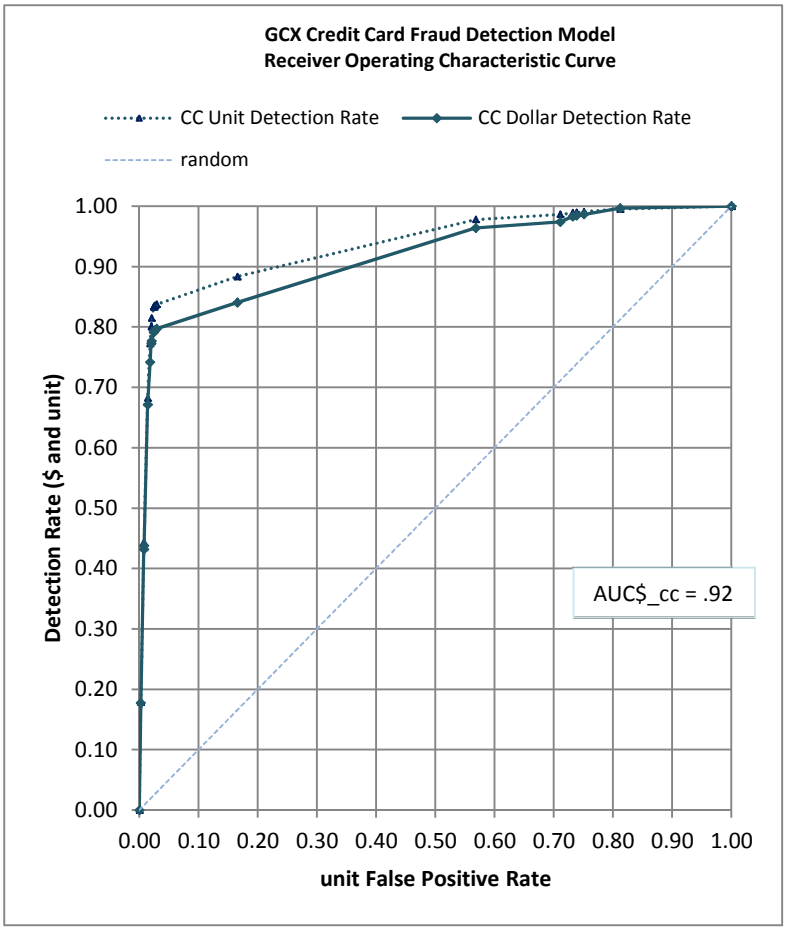

Figure 1. ROC Curve of a GCX Credit Card Fraud Detection Model

THE *CONTINGENCY TABLE* OR *CONFUSION MATRIX*

Fraud detection models are what are more generally known as binary classifiers. A binary classifier assigns an example of a population to one of two categories, and either gets it right or wrong. This process generates a *contingency table* or *confusion matrix*. We will use the term contingency table here since it usually more familiar to people.

The population of transactions or customers to be classified are in one two classes, "Positives," i.e. those transactions that are fraudulent or possibly customers who have been victimized, and "Negatives" i.e. good transactions or customers that are not victims.

[2] This would be the case where something is more often wrong than right, and so one should always do the opposite of what it says.

The detection model will assign an example to either the Positive or Negative (P or N) classes, but it will not do so perfectly. This generates the four possible outcomes in the contingency table. "True" means a correct classification, and "False" means an incorrect classification.

True Positive (TP) – A correctly identified fraudulent transaction or victimized customer

False Positive (FP) – A transaction or customer incorrectly identified as a fraudulent transaction or victimized customer

True Negative (TN) – A correctly identified good transaction or non-victim customer

False Negative (TN) – A transaction or customer incorrectly identified as a good transaction or non-victim customer

These outcomes add up as shown in Table 1 below:

Table 1. Contingency table Terminology (units)

	Frauds	Not Frauds	Totals
Alerted as Fraud or Victim	TP	FP	TP+FP = P
Not Alerted	FN	TN	FN+TN = N
Totals	TP+FN	FP+TN	TP+FP+TN+FN

Table 2 below shows the contingency table for fraud detection system in terms of a monetary basis (e.g dollars):

Table 2. Contingency table Terminology ($)

	Fraud Dollars	Good Dollars	Totals
Alerts	TP$	FP$	TP+FP = P$
Not Alerted	FN$	TN$	FN+TN = N$
Totals	(TP+FN)$	(FP+TN)$	(TP+FP+TN+FN)$

CLASSIFICATION PERFORMANCE MEASURES

Evaluation metrics for binary classifiers derived from the contingency table include the following, and note that these can be computed on a unit basis or a dollar basis, or combinations of units and dollars (e.g. dollar detection rate, or DDR, is TPR based on dollars).

Table 3. Evaluation Metrics for Fraud Detection

Misclassification Rate

$$MR = \frac{FN + FP}{TP + FN + FP + TN} \tag{5}$$

Accuracy

$$A = 1 - MR = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

True Positive Rate

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

False Positive Rate

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

Specificity

$$SPC = 1 - FPR = \frac{TN}{FP + TN} \tag{9}$$

Positive Predictive Value

$$PPV = \frac{TP}{TP + FP} \tag{10}$$

Negative Predictive Value

$$NPV = \frac{TN}{TN + FN} \tag{11}$$

False Discovery Rate

$$FDR = \frac{FP}{FP + TP} \tag{12}$$

False Discovery Ratio[3]

$$FDRatio = \frac{FP}{TP} \tag{13}$$

Alerting Rate or Positive Rate[4]

$$Arate = \frac{FP + TP}{FP + TP + FN + TN} \tag{14}$$

Scoring Classifiers. For detection models that produce scores instead of simply "Positive" or "Negative" classifications, a threshold score is set below which examples are considered "Negative" and above which they are considered "Positive." So all of the measures above become parametric in the alert score threshold value. The contingency table then takes on the extra dimension of *score threshold* or *score bands*. This permits analysis of the operational effectiveness of the classifier over a range of thresholds the produce a feasible number of alerts for the fraud investigation staff. The score-band table also provides a very good and practical approximation of the ROC curve of the classifier.

Non-Scoring Classifiers. Some detection models do not produce a score, just a classification. Rules are a good example of non-scoring classifiers. Because the score is missing, one cannot choose or optimize an operational threshold. Even if a rule system (or *policy*) produces a score, the score does not map or transform coherently into a probability distribution function (PDF); the ROC for such scores will often be disjoint or discontinuous.

ECONOMIC EFFECTIVENESS EVALUATIONS

Two methods of evaluating the economic effectiveness of a detection system are presented here. One uses a fixed operating point and the corresponding contingency table. The other exploits the ROC curve and the associated payoff to determine an economically optimal operating point.

[3] This metric is not discussed much in the classifier literature, since its application and characteristics are not well understood; however it is of practical interest to operational managers in determining the general quality of fraud alerts in a simple and understandable way.

[4] This measure is more useful for operational analysis, as it shows the number of alerts that will be generated by the detection model, and which should be adjudicated by a fraud investigator.

CONTINGENCY TABLE METHOD OF EVALUATION OF ECONOMIC BENEFITS

For a single threshold value partitioning the alerts into the sets for review and set to ignore, a contingency table may be constructed. This is specific to a single threshold value, and not as general as the ROC method. However, when the threshold corresponds to the optimal value, this method is applicable.

The following tables demonstrate the economic effectiveness analysis for a (fictitious) test result on a bank transaction.

Table 4. Cost Factors (Example)

Average Loss	$ 1,400.00
Time to Review Alert (minutes)	12
$/Minute of Reviewer Time	$ 0.76

Table 5. Payoff/Cost Matrix

-Review Cost (TP)+Expected Loss[5]	-Review Cost (FP)
-Expected Loss (FN)	Zero (TN)

Table 6. Payoff Matrix - example

$1,390.88	($9.12)
($1,400.00)	$0.00

Table 7. Contingency Table for System (sample)

	Fraud	Not Fraud	Totals
Alert	149	1,528	1,677
No Alert	69	4,998,254	4,998,323
Totals	218	4,999,782	5,000,000

Table 8. Benefits and Costs (Table 6 × Table 7)

	Fraud	Not Fraud	Totals
Alert	$207,241.12	($13,935.36)	$193,305.76
No Alert	($96,600.00)	$0.00	($96,600.00)
Totals	$110,641.12	($13,935.36)	**$96,705.76**

OPTIMIZING ECONOMIC EFFECTIVENESS ANALYSIS BASED ON THE ROC CURVE

When economic factors are available to assign costs to the various outcomes in the contingency table, we can determine the economic effectiveness of a classifier. This is the case in fraud detection, since we have expected values for a fraud loss, as well as the cost of reviewing the alert. Other cost factors could be introduced, such as the I/T Total Cost of Ownership (TCO) of the computer system doing the fraud detection, but to keep matters simple, we will use the expected loss by transaction type, and the cost per hour of a fraud detection banker. These costs and benefits comprise a payoff matrix for each outcome in the contingency table.

Economically Optimal Threshold. For a given fraud detection system we can vary the threshold below which alerts will not be reviewed by a detection banker. Analysis of dynamic variation in the threshold is not in

[5] This is the value of the loss prevented, i.e. positive benefit

the scope of this analysis (e.g. working alerts down according to available labor as opposed to alert score). For any threshold, we can compute a true positive rate and a false positive rate based on the *ex post facto* analysis of actual frauds versus detected frauds and alerts worked and not worked. A retrospective analysis of this data will produce the ROC curve for the detection model. This then provides, for every threshold value, a net cost (benefit) in terms of dollars. This curve will have a minimum somewhere in the threshold range, and this identifies economically optimal threshold for that detection model.

Comparative Evaluation of Effectiveness. When we have two detection models to compare, we simply identify the economically optimal threshold value for each model, and the model that has greater economic benefit is preferred.

Example of Economic Effectiveness and Optimal Threshold

The results of a hypothetical economic analysis of a fraud detection system are shown here to illustrate the procedure. Scores in this example detection system are integers in [0,100], and so for each score, the threshold for review is set, and the costs of reviewing the alerts, the cost of fraud losses incurred because alerts were not reviewed (below threshold), and the benefit of preempting a loss because the alert was generated and reviewed, and the cost of losses incurred as escapes are computed. This generates a unit cost curve as shown in Figure 2 below.

In this example, alerts scores are normally distributed, Normal(μ=50,σ=20), the expected loss on a fraud is $500, an alert costs $15 to review, and the base rate of occurrence of fraud is .04.

It is a straightforward matter to identify the optimal review threshold by visual inspection of this graph, the optimal threshold, **`Review_Threshold`***, is 35 points.

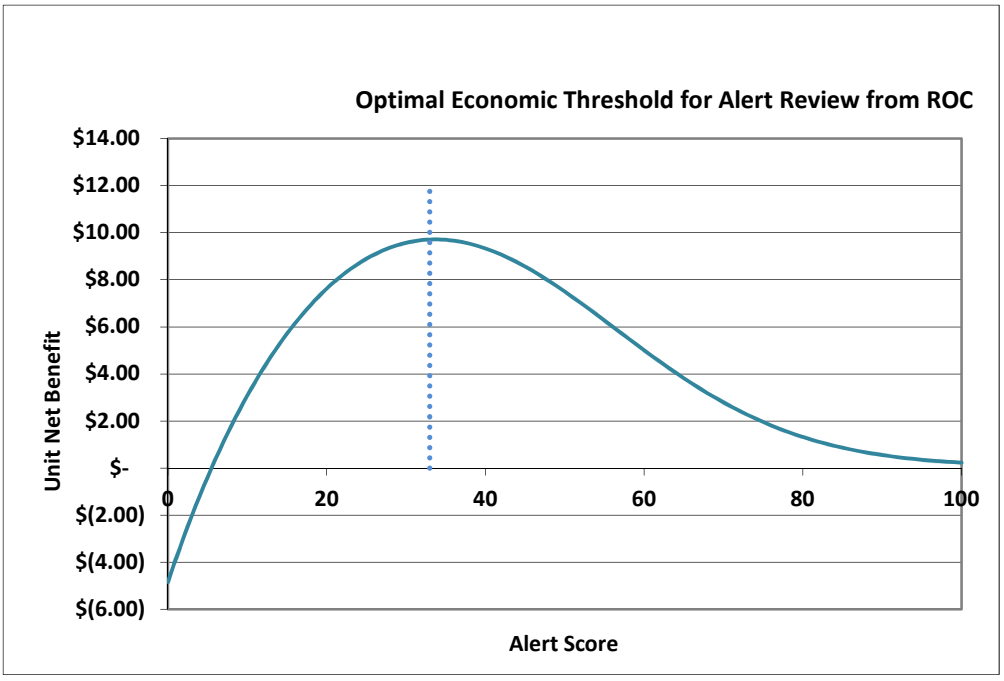

Figure 2. Optimal Score Threshold for Review Example[6]

[6]Source: http://www.answers.com/topic/receiver-operating-characteristic?cat=technology

References

[1] Tom Fawcett, ROC *Graphs: Notes and Practical Considerations for Researchers*, HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto, CA 94304

[2] Tapas Kanungo and Robert M. Haralick , *Receiver Operating Characteristic Curves And Optimal Bayesian Operating Points*, Intelligent Systems Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA 98195

[3] Mithat Gonen, *Analyzing Receiver Operating Characteristic Curves with SAS*, SAS Institute, 2007