The issue of classification versus precision rates in comparing classification models

Informs NYC 2015/09.

By Leonardo Auslender
Independent Statistical Consultant
Leonardo.Auslender ‘at’ gmail ‘dot’ com
Setting and topics of discussion.

Binary Target Models, used in DM, Finance, Telco, Clinical, etc.

Target variable usually denoted by Y and accepts two values or labels, “0” (non-event) and “1” (event). Typically, proportion of events << proportion of non-events.

Predictions are usually classified as “events” (“1”) whenever posterior probability > cutoff. Else, non-event (“0”). “Known” cutoff value assumed in this section. Tree models use an implicit cutoff point, usually .5.

Goals:

Model Comparison.

Decision on model deployment.

Model(s) outcome easily portrayed in 2 x 2 table, called Classification or Confusion Table.

2 sets of examples: 1st based on 3 models of Titanic survivors. Each model differs in variables utilized for prediction and 2nd two fraud models. Model details not shown for brevity. No K-S or other measures discussed much.
Fixed – Cutoff Setting
### Classification (confusion) Table
(similar tables and analysis for other than 2 * 2).

<table>
<thead>
<tr>
<th></th>
<th>Pred Non</th>
<th>Pred Event</th>
<th>Total Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-event</strong></td>
<td>A (TN): true negative</td>
<td>B (FP): false positive</td>
<td>A + B</td>
</tr>
<tr>
<td><strong>Event</strong></td>
<td>C (FN): false negative</td>
<td>D (TP): true positive</td>
<td>C + D</td>
</tr>
<tr>
<td><strong>Total Predicted</strong></td>
<td>A + C</td>
<td>B + D</td>
<td>A + B + C + D = Grand Total</td>
</tr>
</tbody>
</table>

Classification (Accuracy) Rate: \(100 \frac{(A + D)}{\text{grand total}}\).

Positive Precision Rate: \(\text{PPR} = 100 \frac{D}{B + D}\).

Sensitivity = Event class-recall- (hit) rate = \(\text{TPR} = 100 \frac{D}{C + D}\)

Specificity = Non-Event classification rate \(\text{TNR} = 100 \frac{A}{A + B}\).

\(1 - \text{Specificity}\) = Event miscl. rate (false alarm) \(\text{FPR} = 100 \frac{B}{A + B}\).

\(F_1\)-measure = \(\frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}\) (geom. Mean of prec. & recall), (Van Rijsbergen, 1979).

Mutual Information = \(\sum_{t=1}^{2} \sum_{p=1}^{2} P(Y_t, X_p) \log \frac{P(Y_t, X_p)}{P(Y_t)P(X_p)}\)
Classification rate

Proportion of events predicted as events (similarly for non-events). (Also called recall or True Positive Rate (TPR). When considering events and non-events together, called accuracy or overall classification rate of events and non-events).

Denominator is number of original events and non-events, which is fixed for any original ratio of events and non-events (ratio called prior, skew, etc). Thus, in previous slide, TPR = 100 * D / (C + D).

Demsar (2006) shows most algorithms compared based on accuracy.

Precision Rate:

Proportion of predicted non-events/events that are truly non-events/events. Thus, PPR = 100 * D / (B + D).

Denominator is number of predicted events and non-events, which is conditional on any prior number of events.
<table>
<thead>
<tr>
<th></th>
<th>Pred Non</th>
<th>Pred Event</th>
<th>Total Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-event</td>
<td>A</td>
<td>B</td>
<td>E</td>
</tr>
<tr>
<td>A / E = TNR</td>
<td>A / G = NPR</td>
<td>B / E = FPR</td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>C</td>
<td>D</td>
<td>F</td>
</tr>
<tr>
<td>C / F = FNR</td>
<td>D / F = TPR</td>
<td>D / H = PPR</td>
<td></td>
</tr>
<tr>
<td>Total Predicted</td>
<td>G</td>
<td>H</td>
<td>I</td>
</tr>
</tbody>
</table>
Graphical Appreciation.

Predicted Events: everywhere else outside of oval, is predicted non-events.

\[
\text{Pos. Precision} = PPR = \frac{\mid \text{Predicted Events} \cap \text{Event} \mid}{\mid \text{Predicted Event} \mid} = \frac{TP}{TP + FP}
\]

\[
\text{Pos. Classif} = TPR = \frac{\mid \text{Predicted Event} \cap \text{Event} \mid}{\mid \text{Event} \mid} = \frac{TP}{TP + FN}
\]
Example of Accuracy vs. Precision.

Hand (2007, pp. 22-23), fraud study, classifier:
Correctly identifies 99 in 100 legitimate transactions → TNR = 99%
Correctly identifies 99 in 100 fraudulent transactions → TPR = 99%

Assume 1 in 1000 is fraud (prior = prevalence = .001).

<table>
<thead>
<tr>
<th></th>
<th>Pred Legit</th>
<th>Pred Fraud</th>
<th>Total Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legit</td>
<td>99% / 989.01</td>
<td>1% / 9.99</td>
<td>100% / 999</td>
</tr>
<tr>
<td>Fraud</td>
<td>1% / 0.01</td>
<td>99% / 0.99</td>
<td>100% / 1</td>
</tr>
<tr>
<td>Total Predicted</td>
<td>100% / 989.02</td>
<td>100% / 10.98</td>
<td>100% / 1000</td>
</tr>
</tbody>
</table>

**Fraud Precision rate**: 0.99 * 100 / 10.98 = 9.016% → 90% predicted frauds are legitimate transactions.  **Legit Precision Rate** = 989.01 * 100 / (989.01 + 0.01) = 99.9%, almost all predicted legits are legits.
On what information should we focus?

TPR = Pr (Pred(Y) = 1 / Y = 1)

PPR = Pr (Y = 1 / Pred(Y) = 1)

Or the square root of the sum divided by ‘e’?

Pr (breast cancer at 40) = Prior = 1%.
Pr (Positive diagnosis / Cancer) = TPR = 80%.
Pr (Positive diagnosis / no cancer) = FPR = 10%.

40 year old woman had positive diagnosis.

Q.: Pr (Cancer / positive diagnosis)? 80%?, 8% (TPR * PRIOR)?, some fudge of TPR and FPR? (note that TPR + FPR ≠ 100%).

<table>
<thead>
<tr>
<th></th>
<th>Pred Healthy</th>
<th>Pred Cancer</th>
<th>Total Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>90% / 891</td>
<td>10% / 99 (FPR)</td>
<td>99% / 990</td>
</tr>
<tr>
<td>Cancer</td>
<td>20% / 2</td>
<td>80% / 8 (TPR)</td>
<td>1% / 10</td>
</tr>
<tr>
<td>Total Predicted</td>
<td>100% / 893</td>
<td>100% / 107</td>
<td>100% / 1000</td>
</tr>
</tbody>
</table>
Detour: Bayes Theorem:

\[
\text{Bayes Theorem:}
\]

\[
\Pr(C / +) = \frac{\Pr(C) \times \Pr(+ / C)}{\Pr(C) \times \Pr(+ / C) + \Pr(\overline{C}) \times \Pr(+ / \overline{C})}
\]

\[
= \frac{\text{Prior} \times \text{TPR}}{\text{Prior} \times \text{TPR} + (1-\text{Prior}) \times \text{FPR}}
\]

\[
= \frac{.01 \times .80}{.01 \times .80 + .99 \times .10} = .075 = 8 / 107.
\]

Note that in order to find \( \Pr(C/) \), the only necessary information is TPR and FPR that originate from the ‘predicted’ column only.

In Business environment, \( \Pr(C/) \) == \( \Pr(\text{Fraud / predicted fraud}) \)== \( \Pr(\text{Responder / predicted responder}) \). \( \Pr(\text{C}) = .01 \) is no longer relevant for the patient. TPR and FPR ‘known’ from previous work and chosen cutoff.
Bayesian Formula and Modeling.

Bayes’ formula updates original unconditional probability of cancer 1% with the new information from test diagnostic.

Probability jumps from 1% to 7.5% ➔

1) Modeling by “conditioning” on test diagnostic ➔ update of information.

2) Notice that Bayes’ result is ‘our’ measure of precision ➔ once Test diagnostic (model) is available, unconditional probability (PRIOR), TPR, FNR irrelevant (if you believe in given test). PPR and NPR become relevant.
Neglecting Costs and Getting a Cost headache.

If misclassification costs were equal (FNR and FPR costs), then could replace FNR and FPR by overall misclassification measures.

But, cost of classifying as Legit when is actually fraud case VERY DIFFERENT from classifying as fraud when legit.

If classify as Legit when it is a fraud case, business incurs direct loss.

If classify as fraud when legit, business does not directly incur loss but may lose customers due to loss of customer good-will.

Modeler may decide to minimize loss function, or better, maximize profit function, which requires FNR and FPR costs:

\[
\begin{align*}
\text{Loss} &= \pi_0 c_0 FNR + \pi_1 c_1 FPR \\
\pi_i &= \text{Prior Prob (Y}=i\text{), } c_i = \text{miscl. cost into i, } r_i = \text{ith revenue.} \\
\text{Profit} &= \pi_0 (r_0 TNR - c_0 FNR) + \pi_1 (r_1 TPR - c_1 FPR)
\end{align*}
\]
More terminology, just to summarize.
Conditional probabilities.

TPR = Prob (PRED POS / POSITIVE (EVENT)) = TP / (TP + FN) = Cum % captured events (DM).

TNR = Prob (PRED NEG / NEGATIVE (NO EVENT)) = Correct Rejection

FPR = Prob (PRED POS / NEGATIVE (NO EVENT)) = 1 − TNR = False Alarm = Type I error

FNR = Prob (PRED NEG / POSITIVE (EVENT)) = Miss = Type II error

PPR = Positive Precision rate = Purity = Prob (POS / PRED POS)) = TP / (TP + FP) = Cum % events (and from this, lift, etc in DM).

NPR = Negative Precision Rate = P (NEG / PRED (NEG)) = TN / (TN + FN)

Unconditional Probabilities.

Prevalence = risk = P (Positive) Used mostly in clinical studies.
Classification table as goodness-of-fit?

1. Good model may classify poorly: Hosmer and Lemeshow (2000, p. 157): model with expected misclassification rate dependent on coefficient slope, not on model fit of logistic regression (unsystematic distances between observed and predicted).

2. Classification done by choosing cutoff point in posterior probability. Well known that classification favors majority group, which is independent of model fit. Thus, if \( P_1 = .49 \) and \( P_2 = .52 \), and cutoff is .5, observations classified into different categories when probabilities very close.

Assumes known and unchanging “natural” class distribution and that error cost of FP = errors FN. Typically favors majority class; but in most applications, cost of misclassifying “1” is higher.
Accuracy can mislead, comparing 2 models (right and left).

Example 1: Assume “1” important. Left Overall Accuracy = 92.5%,
Right overall accuracy = 97.5%, but right model misses all “1”s.

<table>
<thead>
<tr>
<th>Accuracy= 92.5%</th>
<th>Predicted 0</th>
<th>Predicted 1</th>
<th>Accuracy= 97.6%</th>
<th>Predicted 0</th>
<th>Predicted 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual 0</td>
<td>180</td>
<td>15</td>
<td>195</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Actual 1</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>100%</td>
<td>25%</td>
<td></td>
<td>97.5%</td>
<td>?</td>
</tr>
</tbody>
</table>

Example 2: 80% accuracy in two models. If test data set contains more “0”s, right model better. If more “1”s, left model.

<table>
<thead>
<tr>
<th>80% acc.</th>
<th>Predicted 0</th>
<th>Predicted 1</th>
<th>80% acc.</th>
<th>Predicted 0</th>
<th>Predicted 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual 0</td>
<td>40</td>
<td>10</td>
<td>50</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Actual 1</td>
<td>10</td>
<td>40</td>
<td>20</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>80%</td>
<td>80%</td>
<td>71%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>
Classification table as goodness-of-fit?

3. **Models from different samples** cannot be compared based on these tables because predicted probabilities are confounded by the distribution of probabilities in original samples.

4. **Classification error may increase** when highly significant variable (however defined) is added to model. Classification error is insensitive and inefficient measure. Use Brier’s instead (Pencina et al, 2008). Some authors (Provost et al (1998) prefer ROC, while others (Harrell) used to prefer ROC and switched to Brier (to be discussed in the future).

\[
\text{Brier} = \frac{(\hat{Y} - Y)'(\hat{Y} - Y)}{N}, \quad \hat{Y} : \text{posterior}
\]
What is Important?
Classification Costs Specific by Industry and discipline.

<table>
<thead>
<tr>
<th></th>
<th>Pred non-resp</th>
<th>Pred resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-resp</td>
<td>TN</td>
<td>FP</td>
</tr>
<tr>
<td>Resp.</td>
<td>FN</td>
<td>TP</td>
</tr>
</tbody>
</table>

**DMers:** FNR more important than FPR because FN are not targeted and lose potential revenue of fees & purchases on lifetime basis. Cost of reaching FP usually disregarded. DMers should concentrate on minimizing FN errors. Since FNR = 1 − TPR, equivalent to maximizing TPR, Let's denote (reviewed later) shows higher point of TPR. Max Likel minimizes FP + FN usually.

**In fraud case,** minimizing FNR may still leave bad “precision” for fraud cases.
Classification Costs Specific by Industry and discipline.

Suppose QC of toy screw manufacturing, positive = defective. FP may induce costly retooling, FN depends on toy and intended age, often disregarded. If screw is instead critical part of bridge which could fall, FN obviously costlier.

In Information Retrieval (e.g., web search) TPR not so important because if retrieve 5,000 references, ranking of references of top 50 or so more important than overall TPR. Lift (next section).

In Clinical and marketing analyses, event rates can be very low (called prevalence in clinical, or priors). In this case a test may yield high FP rates that are higher than the FN rates.

In this case a small PPR rate yields high FPR even if the TNR is close to 100%.
Are False positives less Important?

Database Technologies (DBT) was hired by Florida to create list of potential people to remove from list of voters (Ayres, 2007, p.137).

DBT matched registered voters to national lists of felons and was required 90% match between voters’ and felons’ names, (rough) birthday and race → many false positives emerged → 57,746 registered voters identified as convicted felons. (felon = positive).

Since higher proportion of convicted felons are black, higher proportion of black people were false positives than whites.

Clearly, not possible to request 100% matching. And no False Negative study ever done.

Highly debatable issue in civil rights.
Variable cutoff Point.
Changing % of Non-events and Events per prob. cutoff
Let's denote posterior probability or any monotonic transformation thereof as $S(Y)$. Thus, for given $s$ in $[0, 1]$, $TPR = Pr(S > s / Y = 1)$. And likewise, $PPR = Pr(S > s / Pred(Y) = 1)$. $S$ and $s$ in $[0, 1]$ and typically $s = [0, .1, .2, ..., 1]$.

Receiver Operating Characteristics (ROC) curve graphs or tabulates pairs of $TPR$ (in the y-axis) and $FPR$ (1 – specificity, in the x-axis) for different levels of $s$ in $S$. That is, $TPR(s)$ vs. $FPR (s)$.

Lift graphs or tabulates TP $s$ vs. Pr $S > s$ (note: PRECISION based measure), and gains chart tables provide information on cumulative percentage of captured responders and of true events at point $s$. Lift measure is usually divided by the overall response rate, i.e., by $TPR (s = 0)$. Lift tables usually presented as part of Gains Tables.

Note that neither ROC nor LIFT recommends cutoff point straightforwardly.
ROC and Cutoff Probabilities.

ROC building based on simulating probability cutoff changes from 0 to 1 (s). Can imagine series of classification tables for varying probability cutoffs and collecting TPR and FPR. Below, two tables for cutoff = 25% and 75% for models 1 through 3.

<table>
<thead>
<tr>
<th>Study #</th>
<th>Event Status</th>
<th>Predicted as Event Value</th>
<th>Predicted as Non-Event Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Event</td>
<td>0.31</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Non-Event</td>
<td>0.63</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>Event</td>
<td>0.13</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Non-Event</td>
<td>0.65</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>Event</td>
<td>0.17</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>Non-Event</td>
<td>0.65</td>
<td>0.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study #</th>
<th>Event Status</th>
<th>Predicted as Event Value</th>
<th>Predicted as Non-Event Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Event</td>
<td>0.60</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Non-Event</td>
<td>0.96</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>Event</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Non-Event</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>Event</td>
<td>0.45</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Different priors, Same AUROC, Different Accuracy, cutoff and Precision.
ROC model selection. Curves can cross. NW direction ➔ better model. B is preferred to C for low FP rates, but C preferred over B later on. A clearly inferior.

“Best” model: max AUROC (area under curve and above 45° line). But max AUROC may not be ‘best’ for specific cost and class distribution.
Area under curve (AUROC or “c” statistic) shows % of pairs of (0 / 1) such that predicted probability for “1” > predicted probability “0” (same as Mann-Whitney U statistic (Bamber, 1975)) or Wilcoxon test of ranks.

(Also, related to Gini = Sommer’s D = 2 AUROC – 1.)

For randomly chosen event / non-event pair, AUROC = Prob (Event > Non-Event).

AUROC can be considered as the average TPR across all possible FPR.
ROC, Lift and Data priors.

**ROC** unaffected by different data priors or cost distributions (except when target variable was created by dichotomizing from continuous variable and probable measurement error, likely in clinical studies (Brenner, Gefeller, 1997)).

**Note that ROC (and AUROC and KS) depend on probability ranking that does not change with priors.** Precision, lift, F-score are affected by changes in priors.

- Same algorithm applied on 2 test data sets with different class balances but same number of events show same ROC curves, but precision-recall (recall = TPR) curve changes.

**REMARK:** Since cutoff (and model) can be chosen based on Precision and TPR, can use **Precision-Recall** curve to combine them that is affected by different priors (presented below).
Gains Table / Chart generation.

1) Table in descending order of cutoff probability, usually shown in deciles (10%), demi-deciles (5%) or vingtiles (20%) intervals of counts of observations.

2) Events expected to bundle at top of table for ‘good’ model.

3) Obtain PPR and derivations thereof, such as lift and cumulatives. (% true events, %captured all events).

4) Interpretation: search for cutoff “decile” (or percentile, etc) based on criterion (profit, specific cumulative response rate, etc.), and consider as “events” observations above cutoff. Note: percentages and counts are of actual observations, not predicted, since posterior probability is used as index in decile creation.
### Gains Table measures.

“Cumulative” accumulate from top of probability to bottom in similar fashion as ROC from left to right.

**Cum % events = Event Precision.** If 1st decile is contacted/targeted, equivalent to predicting all observations as events ➔ **Cum % events in decile = precision** ➔ lift and gain are mere transformations of precision as well ➔

If cutoff (and sometimes model itself) chosen based on Lift, cutoff (and model) actually based on PRECISION (which is not part of ROC).

**Lift:** % Events in decile / Prior % events in pop, and corresponding cumulative.

**Gain:** (% events in decile / Random % events in decile ) – 1 (not very used).

**Cum % captured events = True positive Rate** (belongs in ROC). As cutoff is shifted down the table, Cum % captured = proportion predicted events that are truly events ➔ if cutoff (and model) based on TPR achieved at specific percentile ➔ cutoff (and model) chosen based on classification measure.
<table>
<thead>
<tr>
<th>Percentile</th>
<th>Model Name</th>
<th>Events Rate</th>
<th>Cum Events Rate</th>
<th>% Event Captured</th>
<th>Cum % Events Captured</th>
<th>Lift</th>
<th>Cum Lift</th>
<th>Brier Score * 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TREES</td>
<td>71.708</td>
<td>71.708</td>
<td>35.945</td>
<td>35.945</td>
<td>3.594</td>
<td>3.594</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>20</td>
<td>LOGISTIC_STEPWISE</td>
<td>34.899</td>
<td>48.993</td>
<td>17.494</td>
<td>49.117</td>
<td>1.749</td>
<td>2.456</td>
<td>22.474</td>
</tr>
<tr>
<td></td>
<td>TREES</td>
<td>30.658</td>
<td>51.183</td>
<td>15.367</td>
<td>51.312</td>
<td>1.537</td>
<td>2.566</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td></td>
<td>TREES</td>
<td>23.467</td>
<td>41.944</td>
<td>11.763</td>
<td>63.075</td>
<td>1.176</td>
<td>2.103</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>40</td>
<td>LOGISTIC_STEPWISE</td>
<td>18.960</td>
<td>34.522</td>
<td>9.504</td>
<td>69.218</td>
<td>0.950</td>
<td>1.730</td>
<td>15.362</td>
</tr>
<tr>
<td></td>
<td>TREES</td>
<td>19.920</td>
<td>36.438</td>
<td>9.985</td>
<td>73.060</td>
<td>0.999</td>
<td>1.827</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>50</td>
<td>LOGISTIC_STEPWISE</td>
<td>13.591</td>
<td>30.336</td>
<td>6.812</td>
<td>76.030</td>
<td>0.681</td>
<td>1.521</td>
<td>11.817</td>
</tr>
</tbody>
</table>
Cumulative % Lift, Model 2.
Single misclassified obs causes this drop.
Many situations in which priors and costs are unknown. Could also be case in which priors and costs actually vary, either in time or by subpopulations. E.g., prevalence of certain diseases differs across races or ethnic groups but the individual cannot be properly identified. Certainly prevalence of diseases and other events changes with time. E.g., obesity rates in the US in the past 30 years.

In these cases, ROC and threshold cutoff not so reliable because ROC does not reflect changes in priors.
Determine cutoff point via ROC or Gains chart.
Compromise?: Cutoff: intersection precision curve and TPR: no theory behind.

Positive Precision & TPR, Model 2.
Money Matters: Profits determine cutoff.

Cutoff point selected at top 80\textsuperscript{th} percentile of prob. Distribution.
ROC components and KS to determine cutoff point.

Highest separation point.

Plot sensitivity and specificity (Y) vs. prob. cutoff points (X). Intersection indicates maximum separation (K-S test). No cost specified.
Using ROC with financial information to determine cutoff via ROC - Direct Marketing Application.

Suppose mailing data base 10,000 candidates. Expect 10% response rate ➔ if mail everybody, expect 1,000 responders. Assume budget constraint that allows to mail to just 3,500 ➔

\[
FPR \times 9,000 + TPR \times 1000 = 3,500, \text{ or } TPR = 3.5 - FPR \times 9
\]

From ROC graph, locate pair (FPR, TPR) that satisfies equation, derive cutoff point and contact those above cutoff point.

But this is wrong because TPR + FPR ne 100%. Instead, use PPR and (1 – PPR).
Direct Marketing Flavor
Plus Gainscharts
Information for Models
Comparison.
ROC, model 1 v. 2 v. 3., Precision, Cutoff, Accuracy from top to bottom, + KS Model #.
Note on ROC Comparisons.

Without looking at AUROC, Models 2 and 3 are preferable to 1. But Model 3 has a far shorter range of pairs of TPR and FPR than model 2.

For specific FNR, underlying cutoff points for each models are different (.35, .42, .22 respectively, from left to right). Implicitly, ROC curves have 3rd axis not shown of descending probability ➔ choosing model based on ROC implies specific cutoff for THAT model cannot necessarily be used for another model, either for implementation or to compare classification tables.

There is no single measure or way to select ‘the’ model. A complex model that validates well may not be so robust to changing business/other conditions as simpler and ‘not so good’ model.
Model 2
Model 3

AUROC2 > AUROC3, but ‘3’ has larger % non-events in lowest grouping.

Non-events higher probability than events \( \Rightarrow \) AUROC < 1.
Model 3 finds higher proportion of events and non-events in extremes of score distribution. If interested in targeting very few positives, model 3 is preferable to model 2, despite fact that AUROC 2 > AUROC 3.

But also note that for non-extreme observations, model 3 is not a good discriminator.

K-S does not provide any useful information short of stating that the models differentiate and providing point of maximum difference. Note that those points are not necessarily optimal as cutoffs.

Likewise, ROC per se does not provide any optimal cutoff point.
Cumulative % Lift, Model 1 vs 2 vs 3.
Cutoff set by maximum Cumulative Profit from gains table. If Model 2 selected, cutoff = 0.70.
Fraud Study:
2 models:
Logistic
And
Trees.
<table>
<thead>
<tr>
<th>Fraudulent Activity yes/no</th>
<th>Model Name</th>
<th>Predicted Class</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>Overall</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Class Rate</td>
<td>Prec Rate</td>
<td>Class Rate</td>
<td>Prec Rate</td>
<td>Class Rate</td>
</tr>
<tr>
<td>0</td>
<td>LOGISTIC_STEPWISE</td>
<td>97.30</td>
<td>83.73</td>
<td>-2.70</td>
<td>-31.01</td>
<td>97.30</td>
</tr>
<tr>
<td></td>
<td>TREES</td>
<td>98.87</td>
<td>84.66</td>
<td>-1.13</td>
<td>-13.92</td>
<td>98.87</td>
</tr>
<tr>
<td>1</td>
<td>LOGISTIC_STEPWISE</td>
<td>-75.86</td>
<td>-16.27</td>
<td>24.14</td>
<td>68.99</td>
<td>24.14</td>
</tr>
<tr>
<td></td>
<td>TREES</td>
<td>-71.91</td>
<td>-15.34</td>
<td>28.09</td>
<td>86.08</td>
<td>28.09</td>
</tr>
<tr>
<td>Overall</td>
<td>LOGISTIC_STEPWISE</td>
<td>83.73</td>
<td></td>
<td>68.99</td>
<td>82.70</td>
<td>82.70</td>
</tr>
<tr>
<td></td>
<td>TREES</td>
<td>84.66</td>
<td></td>
<td>86.08</td>
<td>84.75</td>
<td>84.75</td>
</tr>
</tbody>
</table>

Note that TPR is rather low while PPR is quite high (prevalence = 20%).
Fraud Models
Positive (positive slope) & Negative (negative slope) Prec. Rates vs. Probs.
Conclusions
Conclusions:

1) **Precision** contains information conditioned on the new information available, e.g., clinical test, demographics. It answers the question of *probability of response/sick/survival given present information*.

2) **Do classification and precision measures typically rank order well?** Yes, but they don’t indicate good models necessarily.

3) **Are models results reliable by just these measures?** NO.

4) **To-do list:**
   1) Residual analysis (Pardoe, many papers).
   2) Calibrate models.
   3) Average and/or take median of predictions as additional model.
   4) Inference on lift measures to determine whether model results not very different (Jiang, forthcoming in JASA, Ratner (Web)).
The End