02/05/16

DA ROC Cost Curves

ROC and Priors

- TPR and TNR are not dependent on prior probabilities but accuracy is

$$Acc = p(+) * TPR + p(-) * (1-FPR)$$

$p(+)$ = proportion of + in population written as π_i up to now

Can draw lines of iso accuracy on ROC curve

$$TPR = \frac{Acc - p(-)}{p(+)} + \frac{p(-)}{p(+)} * FPR$$

- We are looking for the point on the convex hull where the slope of the tangent $= p(-)/p(+)$
- Operating conditions $p(+) p(-)$
- Difficult to imagine slopes of tangents

A new Space

- Going to plot Error vs $p(+)$
- $Err = (FN - FP) * p(+) + FP$
- New space y-Axis: Error   X-Axis $= p(+)$
- Points in ROC space are going to be mapped onto lines in "error/$p(+)$" space
- Allows us to see how error rate varies according to $p(+)$
- How the model will behave under various conditions
- Easier to see when $p(+) = 0$ and $p(+) = 1$
- Gives points $(0, FP)$, $(1, 1-TP)$ for each TP and FP combination
- Always negative rule (classifying all as negative) $FP = 0 = TP$ $(0,0)(1,1)$ / neg
- Always positive rule $TP = FP = 1$ $(0,1)(1,0)$ \ positive
- look at the lower envelope of the graph

the values between diagonals shows range they operate in

If lines are above diagonals → always negative/positive will be ideal → the model is useless

Costs

- Incorporate cost information
- Total cost $= TP \times C(+|+) + FN \times C(+|-) + FP \times C(-|+) + TN \times C(-|-)$
- Objective: minimise cost
- $C(+|-)$ cost of misclassifying + as −, vice versa
- Assume $C(+|+), C(-|-) = 0$
- Can incorporate costs into growing and pruning the tree as well

Growing Tree using cost information

- Can use/not use cost in growing a tree
- GINI without cost: $g(t) = \sum_{i=1} \sum_{j \neq i} p(j|t)p(i|t)$
- GINI with cost: $\sum \sum c(j|i) p(i|t) p(j|t)$ (costs j as i)
- Without costs $r(t) = 1 - \max p(j|t)$ (node level)
- With costs $r(t) = \min \sum C(j|i) p(j|t)$ (node level)
- $R(T)$ (misclassification of tree) $= \sum_{|T|} r(t)p(t)$ (tree level)
- Therefore costs do alter pruning
- Difficulty to estimate cost.
- Ratio of costs is important.

Incorporating Costs into Model

- $Err = p(+) \times FN + p(-) \times FP$
- $Cost = p(+) \times FN(+|-) + p(-) \times FP(-|+)$ = Expected cost ($Ecost$)
- Define our performance line in terms of costs and prior
- Maximum value of Ecost: All cases incorrectly classified.
  $= p(+) \times c(+|-) + p(-) \times c(-|+)$
- Norm(Ecost) = ECOST/MAX COST
- Now plot Norm Ecost vs a function of $p(+)$ and costs

05/16

DA : COSTS

- p(+) is also redefined to include misclassification costs
- Multiply p(+) by C(+|-) and normalize it so can go from 0 → 1

$$PC(+) = \frac{p(+) * C(+|-)}{p(+) * C(+|-) + p(-) * C(-|+)}$$

- For equal misclassification costs $PC(+) = p(+)$

- $PC(+) = 0$ when $p(+) = 0$ or $C(+|-) = 0$
  $PC(+) = 1$ when $p(-) = 0$ or $C(-|+) = 0$

- Expected normalized cost Norm E Cost = $FN * PC(+) + FP * PC(-)$

Recall: ratio of number of relevant records retrieved to the total number of relevant records in the database - measures how well a search system finds what you want

Precision ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved - how well it weeds out what you don't want.

13/04/16 DA

MODEL EVALUATION

- Results make sense with preliminary analysis?
- Make sense with background of data?
- Splits and terminal node make sense?
- # of terminal nodes
- Test data
- Applied to many different types of models

Test v Training Set:

- Split data into training and test data
- Model built on training data
- Use test set for evaluation
- Typically 80:20 or $\frac{2}{3}$ / $\frac{1}{3}$
- Make sure you have enough cases for each target category - use stratification
- Has to be done Randomly
- Test data should reflect future data

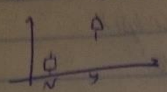
Methods

- Simple plots and summary statistics
- Confusion matrices
- ROC - Receiver Operating curve
- Lift chart
- Role of costs and prior probabilities

- Use test data as input
- Calculate probabilities or values

- Trying to evaluate how close predicted probabilities are to be true probabilities

- In a good model should be a clear distinction between yes and no outcome i.e.



- A bad model will have plots that are basically the same ⇒ no distinction between groups

Briers Score

- N cases, 2 groups
  $\hat{p}_i$ - predicted probability used to assign to group
  t_i - target value 0 or 1
  $= \frac{1}{N} \sum (t_i - \hat{p}_i)^2$
- like mean square error
- Small when probability estimate is small and $t_i = 0$ and probability estimate is big and $t_i = 1$

Create Confusion table.

- Pick a threshold/cutoff based on the predicted probabilities, i.e. 0.5
- If $p_i < 0.5 \Rightarrow 0$ $\quad p_i > 0.5 \Rightarrow 1$
- cut offs can vary
- classify into event and non-event

		Predicted +	Predicted −
Actual	+	TP	FN
	−	FP	TN

Want TP and TN big

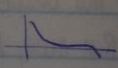
Sensitivity = $\frac{TP}{TP + FN}$ Measure of accuracy for predicting target events (1's)
TPR

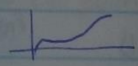
Specificity = $\frac{TN}{TN + FP}$ Measure of accuracy for predicting non target events (0's)

False Positive Rate = 100 - True Negative rate)

% cases correctly classified: $\frac{TP + TN}{TP + TN + FP + FN}$ ← opposite is misclassification rate
Sometimes call accuracy (ACC) or misclassification rate = 100 - ACC

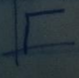
- Could incorporate a cost associated with wrongly classification

- Can create plots such as True positive rate v.s cutoff create 
  - ↳ At 0 cutoff, all positives are correctly classified
  - ↳ At 1 cutoff, all positive are incorrectly classified

- True negative rate vs. cutoff 
  - ↳ At 0 all classified as positive -
  - ↳ At 1 all classified as negative - 100 x rate

- Plot accuracy v.s. cutoff or all graphs together
- Models may have same accuracy but different sensitivity and specificity
- Rare events → the accuracy calculation will be swamped by other event number

ROC Curve

- Plots true positives v.s false positives for a selection of cutoff
- TPR vs FPR = (1 - TNR) for a selection of cutoff
  Sensitivity vs 1 - Specificity
- Calculate at different cutoff
- Good curve (0,0) (0,1) and (1,1) 

- Look at Area Under the Curve AUC (want as close to 1)
- Bootstrap and create a number of ROC curves and calculate a CI
- 45° line for random model
- Check model parameters, do they make sense?
  - calculate predicted probabilities
  - Draw boxplot   look at ROC curves etc

- Sometimes more important to get a high sensitivity or specificity

Model Evaluation - Alternative Approach

- Two outcomes with $N = 3333$ # yes = 483
- Run model, logistic regression or tree
- Produce predicted values using test set P_i (probability of churning)
- Sort the data into 10 parts - deciles

- best case is

Predicted	No	Yes
	0	N
	0	N
	.	.
	N	0
	N	0
	N	0

(all yes values at top, all no values at bottom)

- Cumulative approach - cumulative value in cell downwards and can then calculate % captured

- N = # in sample

N_r = total no. of response where = 1 (defined as an event)

No. in each decile $N/10 = N_d$

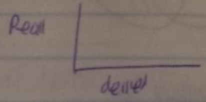
No in cumulative % captured response in decile i : $\frac{N_{ri}}{N_r} \times 100$

cumulative row %'s same as precision

Cumulative: % Captured response ≤ decile i $\frac{\sum N_{kj}}{N_k} \times 100$

% captured response: column %'s for yes "Recall" in R
% Respondents - Row % for yes or "Precision" in R

Lift - $\frac{\text{\% Captured Response}}{\text{\% Random capture}}$



Lift for decile i = $\frac{\text{\% captured response}}{10}$ for non cumulative. - expect 10% in each decile hence 10

Lift for decile i = $\frac{\text{\% captured response}}{\sum_{j=1}^{i} 10}$ For cumulative approach i.e 30 for 3rd decile

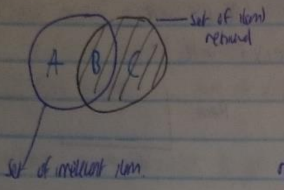
- Want lift high in top deciles then low - how much better than chance model
- % response/precision - want it to decrease to zero quick.

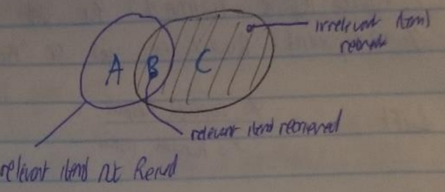
Possible Scenarios
- Can plot for model results, random results and best possible results
- Performance chart in Rattle
  - Adjustment = cumulative % captured response / by % caseload
  - Strike rate = cumulative % Response / by % caseload
  - Black line = cumulative random % captured response

- Confusion matrix
- Misclassification rate
- ROC
- Lift
- Precision
- Accuracy
- Recall
- Cut off

/04/16

Which tool? - Dependent on study
- What you want to optimize





Recall $= \frac{B}{A+B}$

A: number of relevant records not retrieved

B: number of relevant records retrieved

Precision $= \frac{B}{B+C}$

B: # of relevant records retrieved

C: # of irrelevant records retrieved

RECALL - Ratio of number of relevant records retrieved to the number of total # of relevant records in the database

PRECISION: ratio of # of relevant records retrieved to the total number of irrelevant and relevant records retrieved

		model +	model −
actual	Relevant +	TP	FN
	Irrelevant −	FP	TN

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

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

$TP + FP$ = # of items retrieved

AS recall ↑ precision ↓ or recall ↓ Precision ↑

- Categorizing items as relevant or irrelevant
- Determining the # of relevant events in the database
- Recall measures how well a search system finds what you want, precision measures how well it weeds out what you don't want

4/16 DA

CARAT PACKAGE OUTPUT

Accuracy: predicted v actual (reference) values

Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$

95% CI: CI for accuracy CI: $p \pm 1.96 * \sqrt{\frac{p(1-p)}{n}}$ (t-value)

or bootstrap and take percentile

- wide because sample size is small

No Information Rate: How you would do without model, if I predict everything as a yes.

P-Value: How much better you did with model (accuracy) than without it (NIR)

Kappa: First row/ $\frac{\text{actual yes}}{\text{total}}$ agreement controlling for chance

McNemar's Test P Value: paired chi-squared test, suggests no difference in example (probability of 1)

looking at the 6 and 5 in example table

Sensitivity: Measures accuracy of predicting "yes" correctly $\frac{TP}{TP + FN}$

Specificity: Measures accuracy of predicting "no" correctly $\frac{TN}{TN + FP}$

H_0: $Acc \leq NIR$ v H_1: $Acc > NIR$, evidence against H_0, accept alternative

Pos Pred Value: Percentage of predicted yes over all yes

Neg Pred Value: Percentage of predicted no over all nos.

Prevalence: Prevalence of event 27/51 total left column / Total

Detection rate = $\frac{TP}{TP + TN + FN + FP}$

Detection Prevalence = $\frac{TP + FP}{TP + FP + TN + FN}$

Balanced Accuracy = (Sensitivity + Specificity)/2

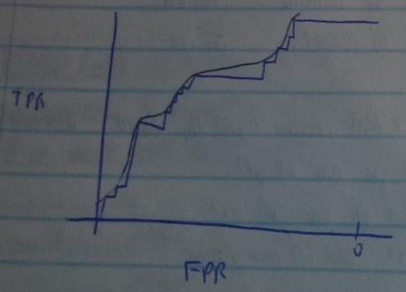
Example of McNemar

B misclass	A misclass: y	A misclass: N
y	n_{00}	n_{10}
N	n_{01}	n_{11}

- Information only in off diagonal elements n_{01} n_{10}
- No difference between the classifiers would expect $\frac{n_{01}+n_{10}}{2}$ in each cell

$$\chi^2_1 = \frac{(O-E)^2}{E} = \frac{\left(n_{01} - \frac{(n_{01}+n_{10})}{2}\right)^2}{\left(\frac{n_{01}+n_{10}}{2}\right)} + \frac{\left(n_{02} - \left(\frac{n_{01}+n_{02}}{2}\right)\right)^2}{\left(\frac{n_{01}+n_{02}}{2}\right)} \quad d.f = 1$$

Reduces to: $\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$ χ^2 with $d.f. = 1$



ROC with concavities and ROC curve convex hull

ROC and PRIORS

- TPR and TNR are not dependent on prior probability but accuracy is:

$Acc = p(+) * TPR + p(-) * (1 - FPR)$

$p(+)$ = proportion of + in population written as π, up to now

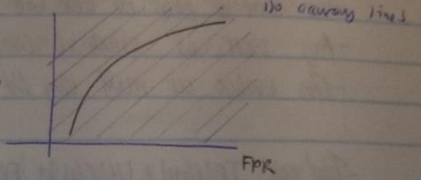
$p(-)$ " of − " "

$p(+) + p(-) = 1$

13/04/16 DA

Can draw lines of ISO accuracy on the ROC curve

$$TPR = \frac{Acc - p(-)}{p(+)} + \frac{p(-)}{p(+)} * FPR$$



Convex Hull

- All points on convex hull dominate
- looking for point on convex hull where slope of tangent = $p(-)/p(+)$
- Operating condition $p(-)/p(+)$

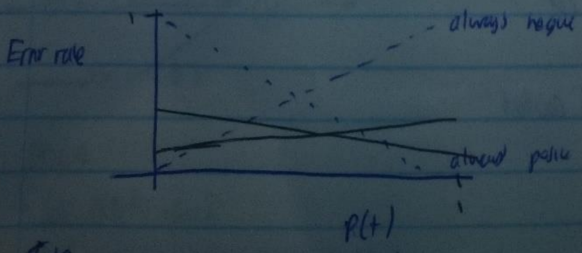
A new Space

- Plot error vs $p(+)$
- $Err = 1 - Acc = 1 - [p(+) * TPR + p(-)(1 - FPR)]$

  $Err = p(+) FNR + p(-) FPR$

  $= (FN - FP) * p(+) + FP$
- New space: y-Axis: Error   x-Axis: $p(+)$

- Points in ROC curve are going to be mapped onto lines in "error/p(+)" space
- Allows us to see clearly how error rate varies according to $p(+)$
- How the model will behave under various conditions.



Take minimum convex hull

- look at the envelope
- look at values for each $p(t)$
- For some of them the best will be always positive or always negative
- May want to attach a cost/profit/loss to each misclassification
- Then evaluate our model using the information

- Total cost: $TP^* C(+|+) + FN^* C(+|-) + FP^* C(-|+) + TN^* C(-|-)$
- Objective to minimise cost
- Can incorporate cost into growing and pruning a tree as well

Growing Trees with Cost

- Gini with cost: $g(t) = \sum_{i=1}^{c} \sum_{j \neq i}^{c} p(j|t)\,p(i|t)$

$$\sum_{i=1} \sum_{j \neq i} C(j|i)\,p(i|t)\,p(j|t)$$ with cost j as i

- Without cost misclassification = $r(t) = 1 - \max_j p(j|t)$ node level
- With cost $r(t) = \min_i \sum_j C(j|i)\,p(j|t)$ node level

$R(T) = \sum_{|T|} r(t)\,p(t)$ → tree level

∴ costs alter pruning regime

Difficulties

- Clunky model of cost
- Difficulty to estimate cost
- What is important is ratio of cost
- Objective changed
- Many more complicated models available