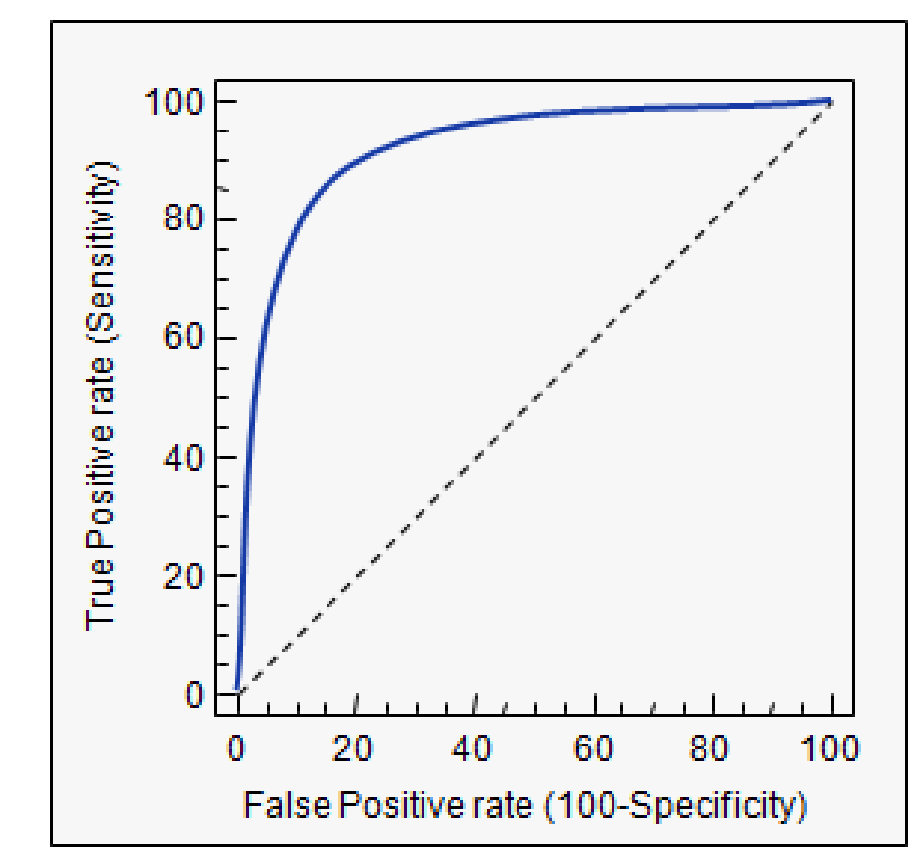
BUSINESS ANALYTICS

Data Mining Techniques

MODEL TESTING

- The receiver operating characteristic (ROC) curve is one of the most useful testing methods for binary classification problems, since it provides a comprehensive and visually attractive way to summarise accuracy of predictions.

ROC CURVE



- The closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.
- The closer the ROC curve is to the 45-degree diagonal the less accurate the test.

RECEIVER OPERATING CHARACTERISTIC CURVE

- The true positive rate (sensitivity)(y-axis) is plotted with the false positive rate (specificity) (x-axis).
- Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.
- As specificity decreases (widen the decision threshold) the sensitivity increases (more sensitive and thus more false positives)

AREA UNDER THE CURVE

- Test accuracy is also shown as the area under the curve.
- The greater the area under the curve, the more accurate the test.
- The diagonal line in a ROC curve represents the perfect chance. In other words, a test that follows the diagonal has no better odds of detecting something than a random flip of a coin. The area would be 0.5
- Area over 0.7 indicate a good model.
- Area over 0.8 indicate a strong model.
- A perfect test has an area under the ROC curve of 1.

AREA UNDER THE CURVE

- This performance metric for binary classification models has two great strengths:
  - The AUC results do not change with changes in the incidents of the actual condition.
  - The AUC is not affected by changes in the relative cost of the two different types of errors (false positive, false negative).
- Therefore, when either future incidents or the cost of classification errors or both are unstable or cannot be known, the AUC is generally the best possible performance metric available.

CALCULATING AOC

- The scores for the observations first must be ranked.
- At each threshold you classify the number of false positives/false negatives (you can construct a confusion matrix for each threshold)
- Next you calculate the false positive rate and true positive rate for each threshold.
- This creates ordered pairs to map on to the ROC curve.
- FP rate = x-axis
- TP rate = y-axis

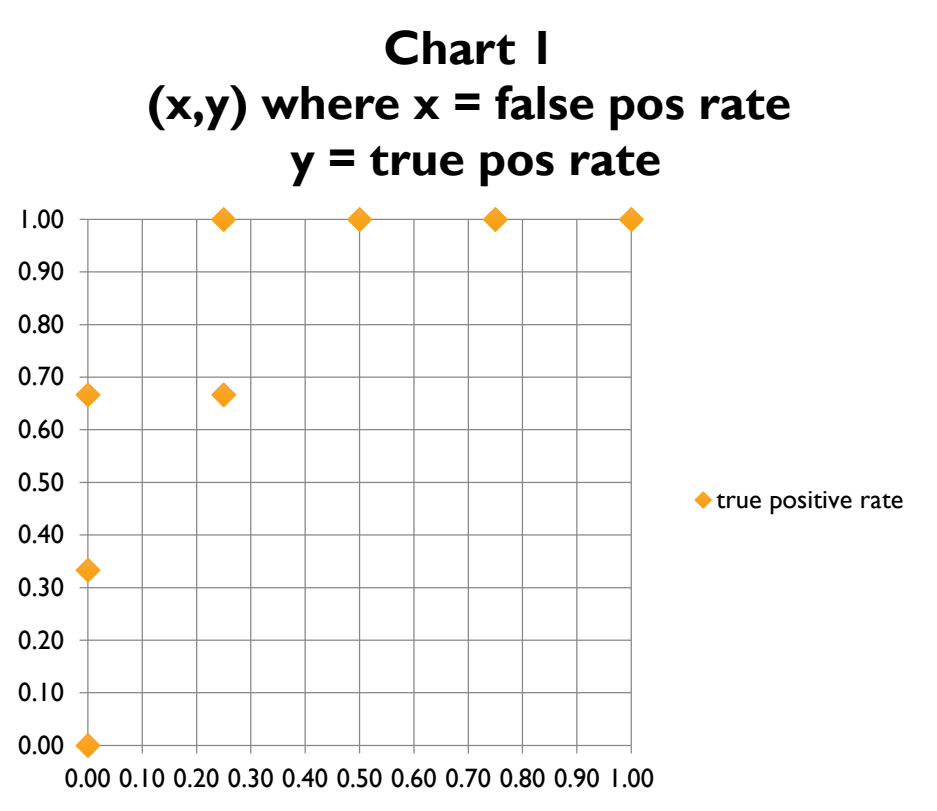
CALCULATING THE FP & TP RATES

- False positive rate is calculated by dividing the number of false positives by the total number of actual negatives.
- True positive rate is calculated by dividing the number of true positives by the total number of positives.

count of events classified as "positive" at each threshold	scores [-3, 3]	outcomes [0 or 1]	false positives at threshold	true positives at threshold
0	[threshold above top score]		0	0
1	3	1	0	1
2	2	1	0	2
3	1	0	1	2
4	0	1	1	3
5	-1	0	2	3
6	-2	0	3	3
7	-3	0	4	3
total number of events:			total number of negative outc	total number of positive outcomes
7			4	3

x axis values equal false positive *rate*	y axis values equal true positive *rate*
0	0
0	0.33
0	0.67
0.25	0.67
0.25	1
0.5	1
0.75	1
1	1
1	0

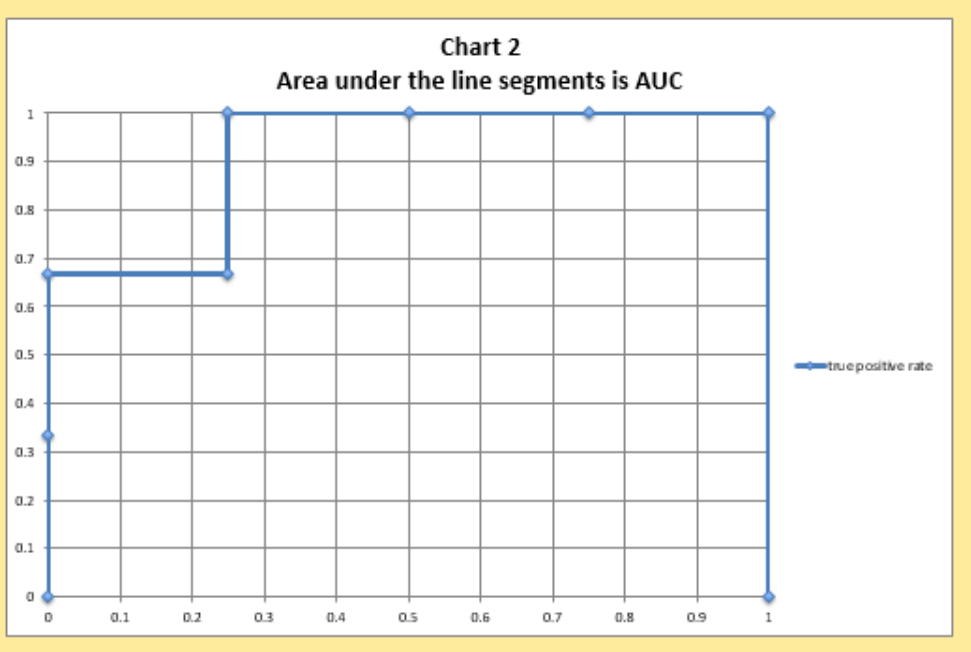
ROC CURVE



- Using the ordered pairs plot the x and y-axis points
- This is done using a scatter chart in excel and you then can set a line marker to draw between the data points.

x axis values equal false positive *rate*	y axis values equal true positive *rate*
0.00	0.00
0.00	0.33
0.00	0.67
0.25	0.67
0.25	1.00
0.50	1.00
0.75	1.00
1.00	1.00
1.00	0.00

AOC



- You then calculate the area under the curve by calculating the area for each rectangle.
- The first here is .25 wide by .67 high thus = 0.1675
- The second is .75 wide by 1 = .75
- The total area = 0.917

- This means this model is considered very good.