Additional classification model techniques

Charles Nicholson

DSA/ISE 5103

Accuracy is not enough

- Accuracy may be an insufficient for evaluating the quality of a classifier – especially in highly imbalanced data
  - e.g., rare disease diagnosis: if only 1% of the population has the disease, then if you ALWAYS predicts "no disease" would have an accurate of 99% → even with no model at all!
- This one reason we have already discussed several other metrics and techniques
  - e.g., kappa, confusion matrices, sensitivity, specificity, positive predictive value, negative predictive value, ROC, AUC, K-S charts, etc.
- We will now add a few more techniques to the list

Remember

		Actual values	
		positive	negative
Predicted values	positive	**TP** (true positive)	**FP** (false positive)
	negative	**FN** (false negative)	**TN** (true negative)

Remember

- **Precision** a.k.a. **Positive Predictive Value**: proportion of predicted positive cases correctly identified

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

- **Recall** a.k.a. **Sensitivity** a.k.a. **True Positive Rate**: proportion of actual positive cases which are correctly identified

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

- **Specificity** a.k.a. **True Negative Rate**: the proportion of actual negative cases which are correctly identified

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

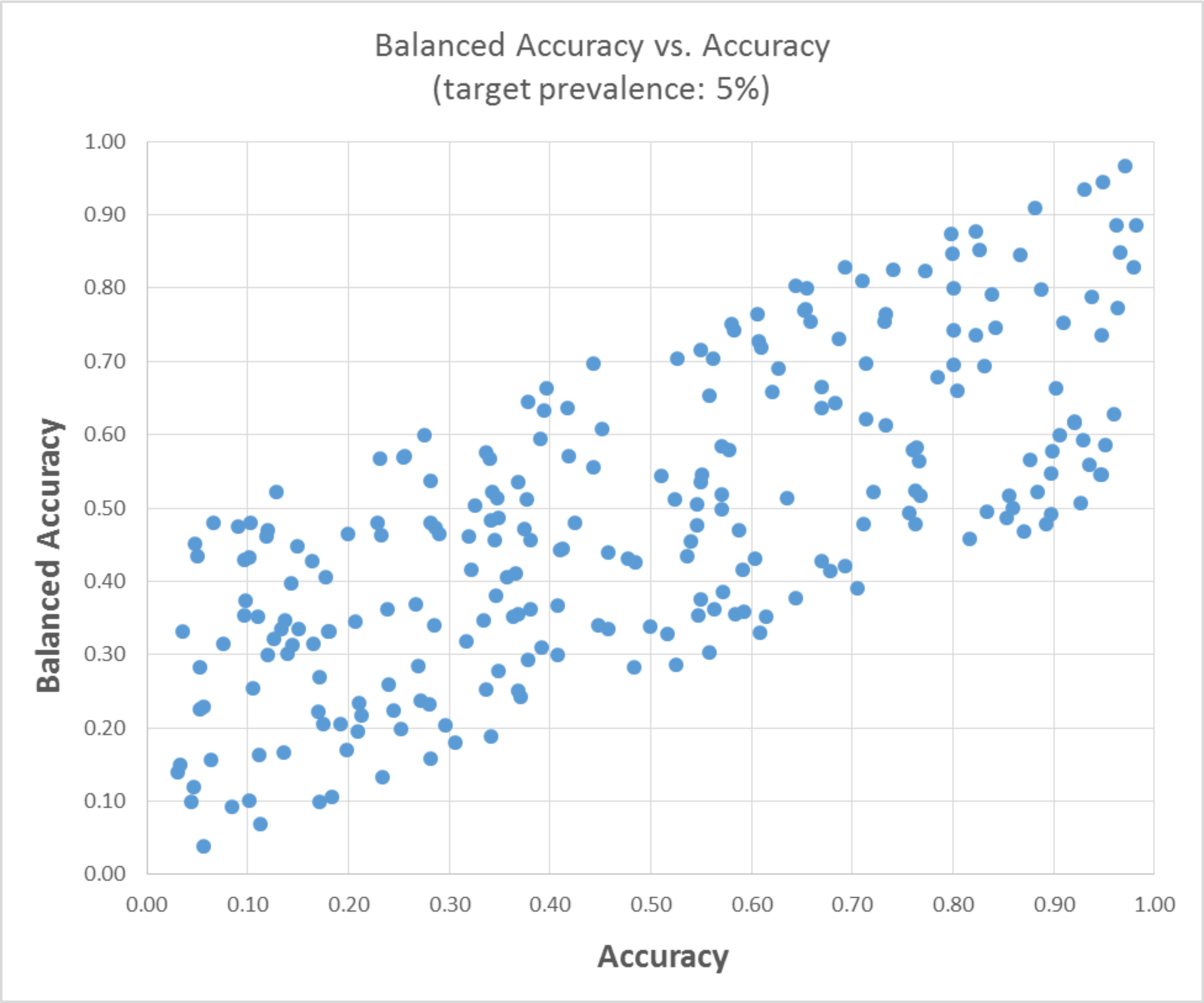
Balanced Accuracy

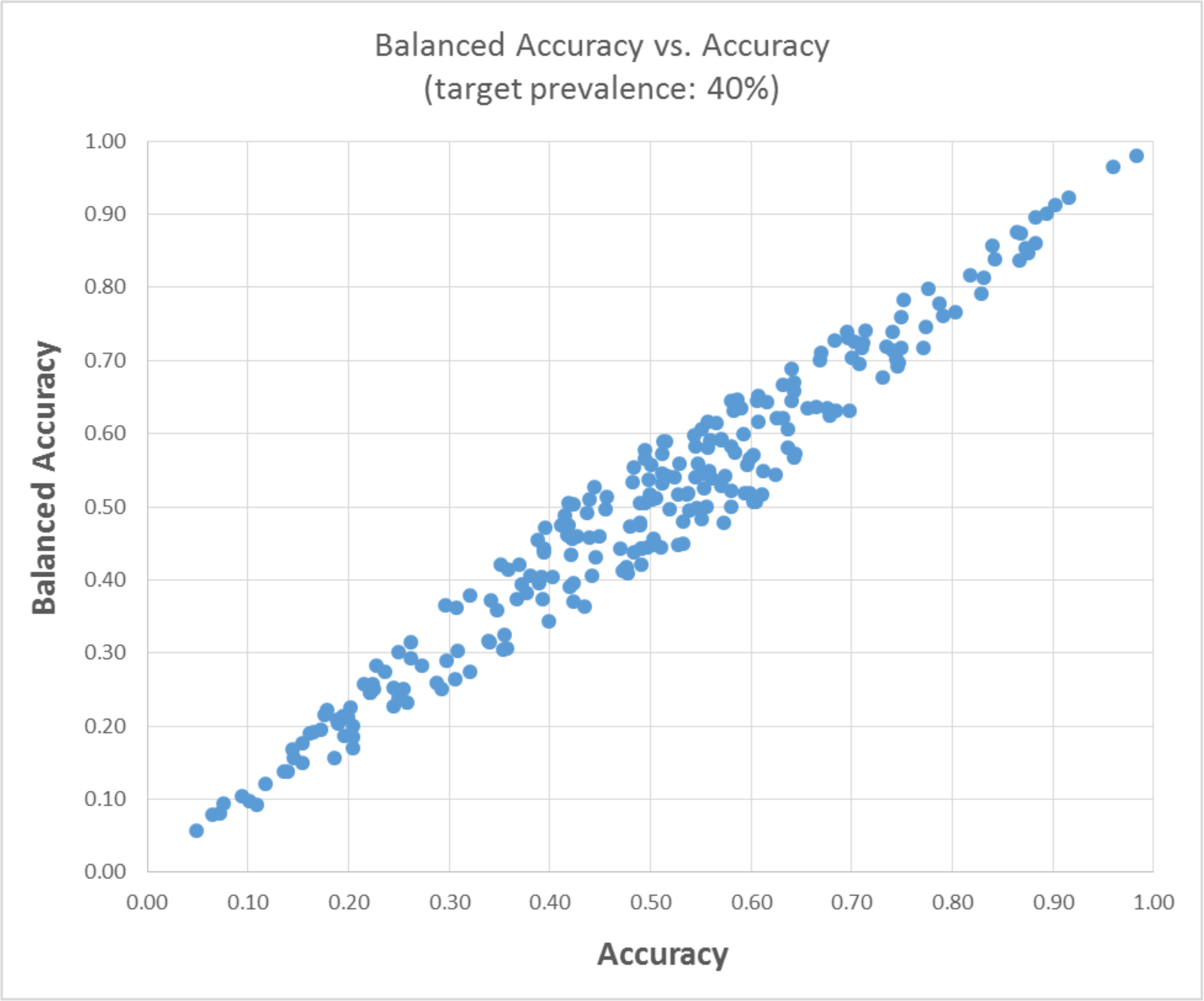
- Average of sensitivity (a.k.a., recall) and specificity

$$\begin{aligned}\text{Balanced accuracy} &= \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right) \\ &= \frac{1}{2}(\text{sensitivity} + \text{specificity})\end{aligned}$$

- For target prevalence equal to 50%, identical to accuracy
- As target prevalence moves away from 50%, balanced accuracy may be very different from accuracy

Balanced Accuracy vs. Accuracy





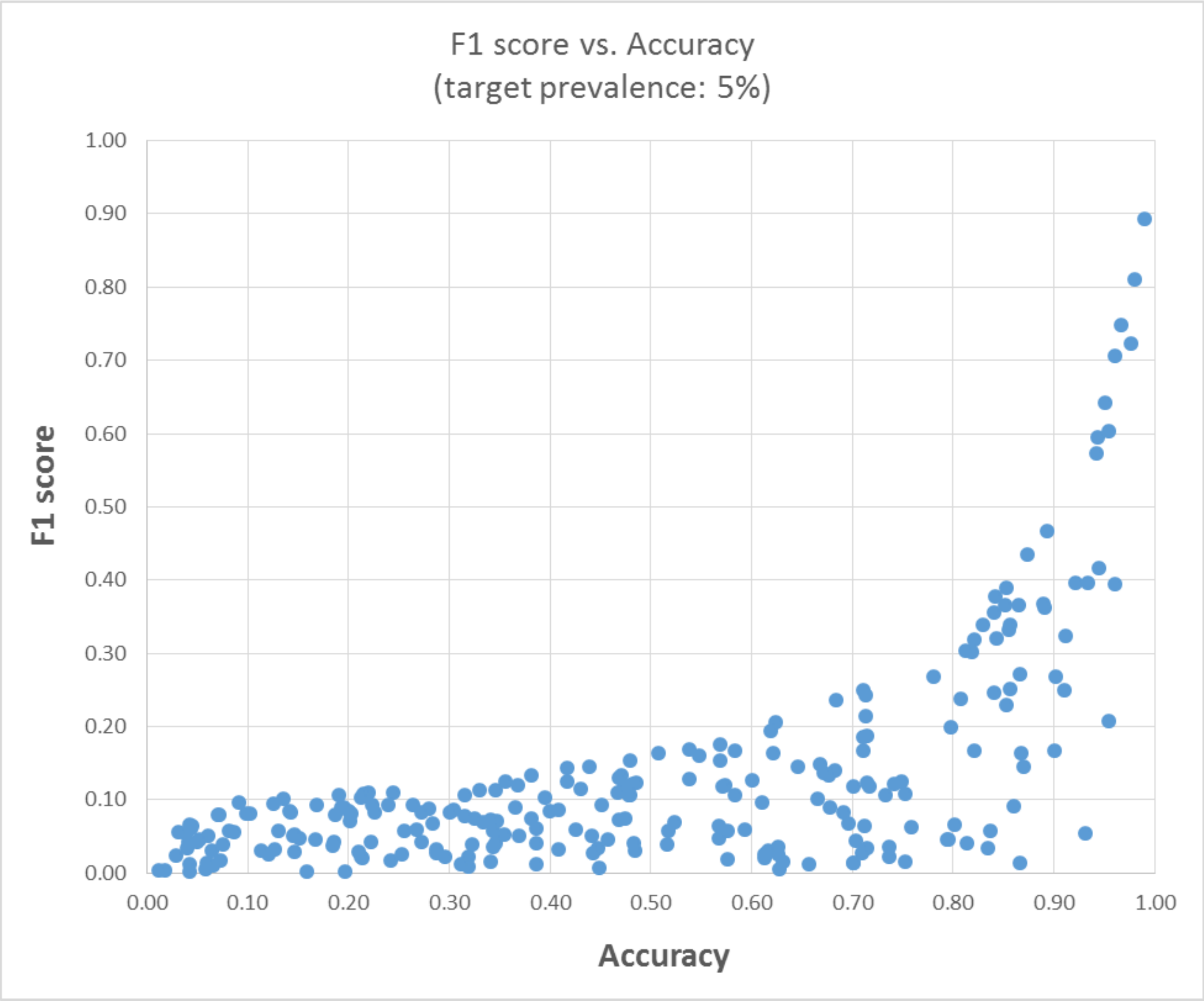
F1 score

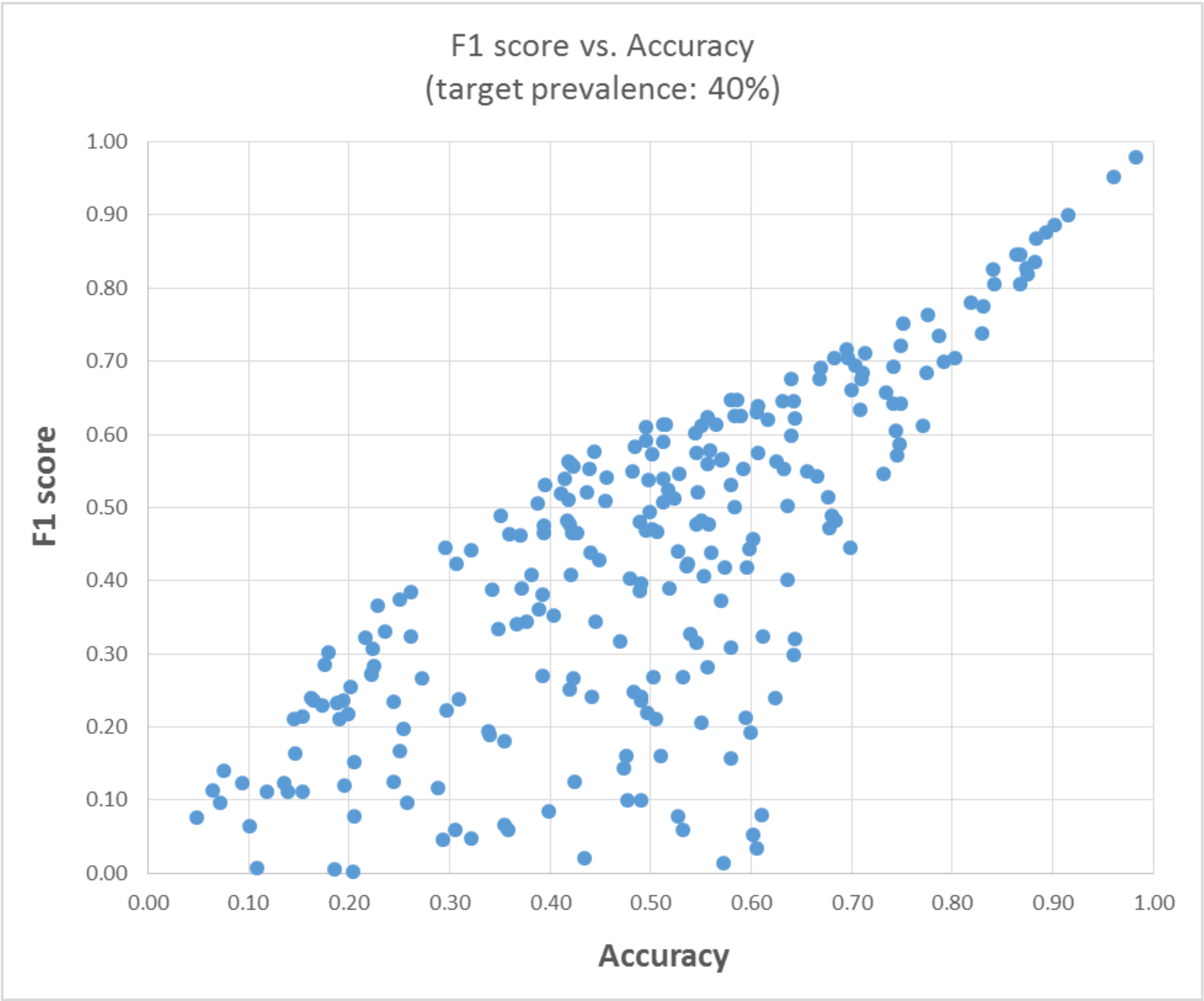
- Harmonic mean of precision and recall

$$\text{F1 score} = 2\left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right)$$

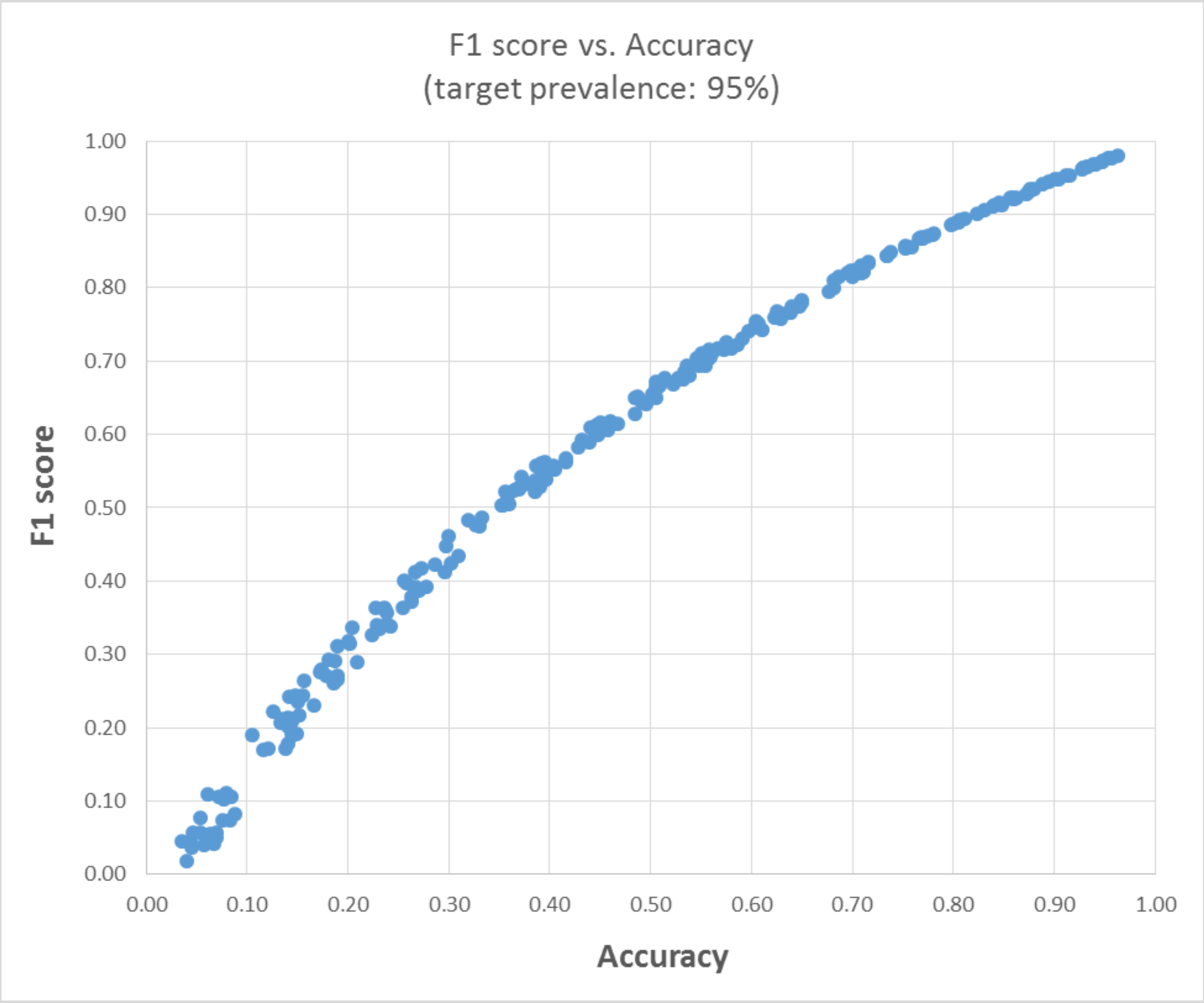
- Also a "balanced" metric; but note it does not consider TN
- Behaves differently based on level of target prevalence
  - For low to medium target prevalence, quite distinct from accuracy
  - For high target prevalence, F1 score similar to accuracy (may be higher than accuracy)

F1 score vs. Accuracy





F1 score vs. Accuracy



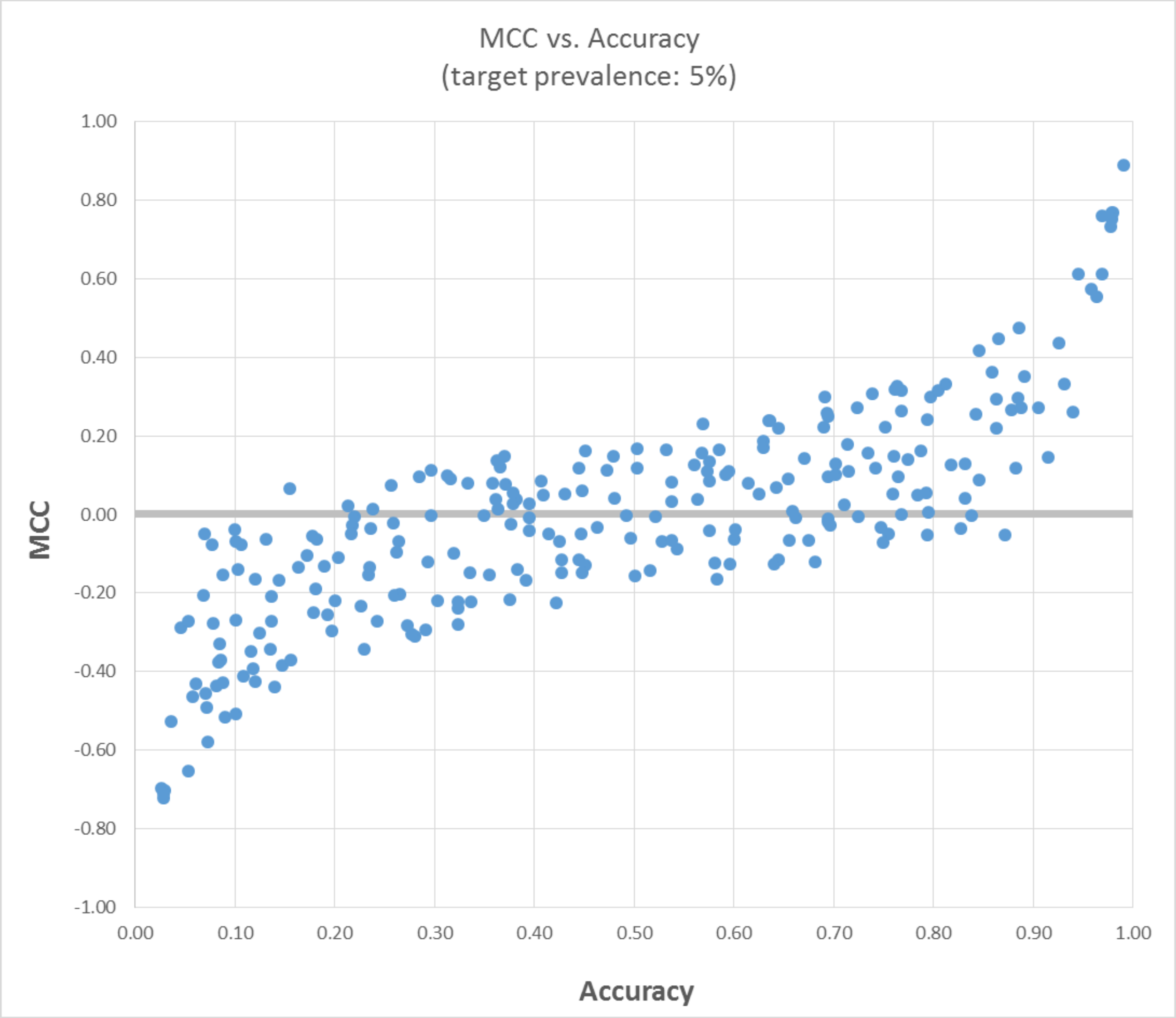
Matthews correlation coefficient

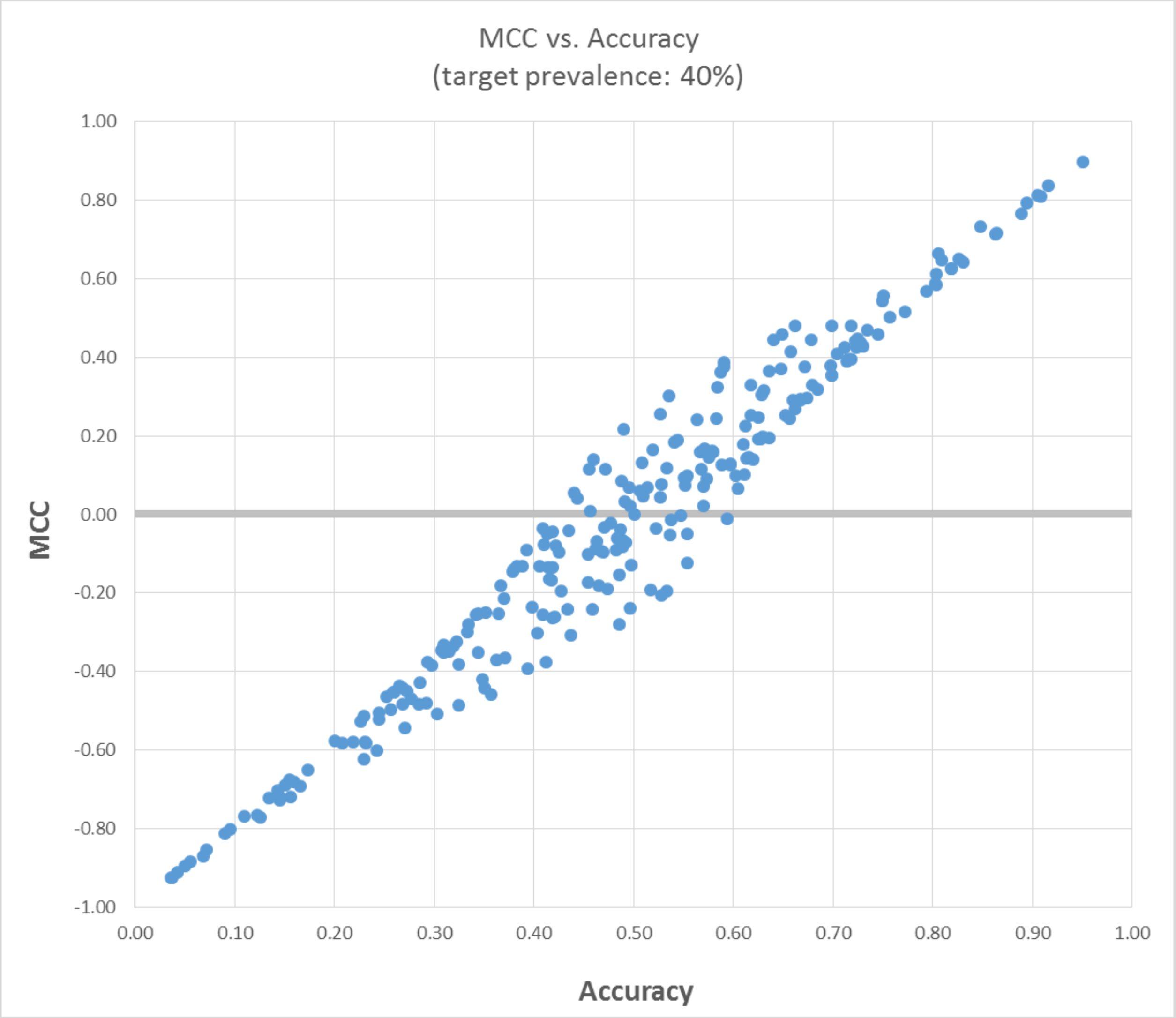
- Another “balanced” measured

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

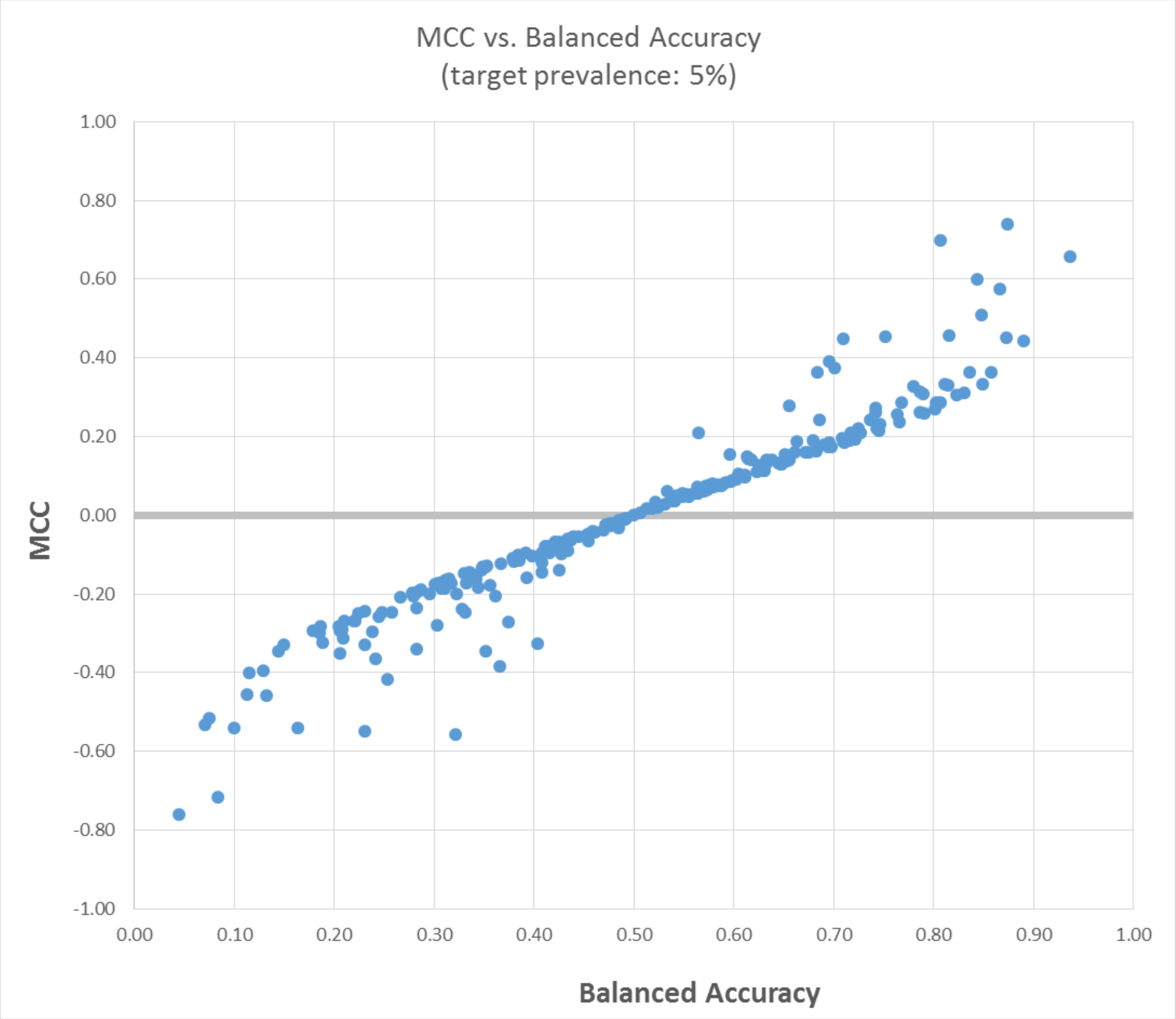
- Takes into account all elements of confusion matrix (unlike F1 score)
- Unlike F1 score, low target prevalence and high target prevalence are have similar impact
- Values are between -1 and +1.
  - +1 is a perfect model and -1 is a perfectly wrong model

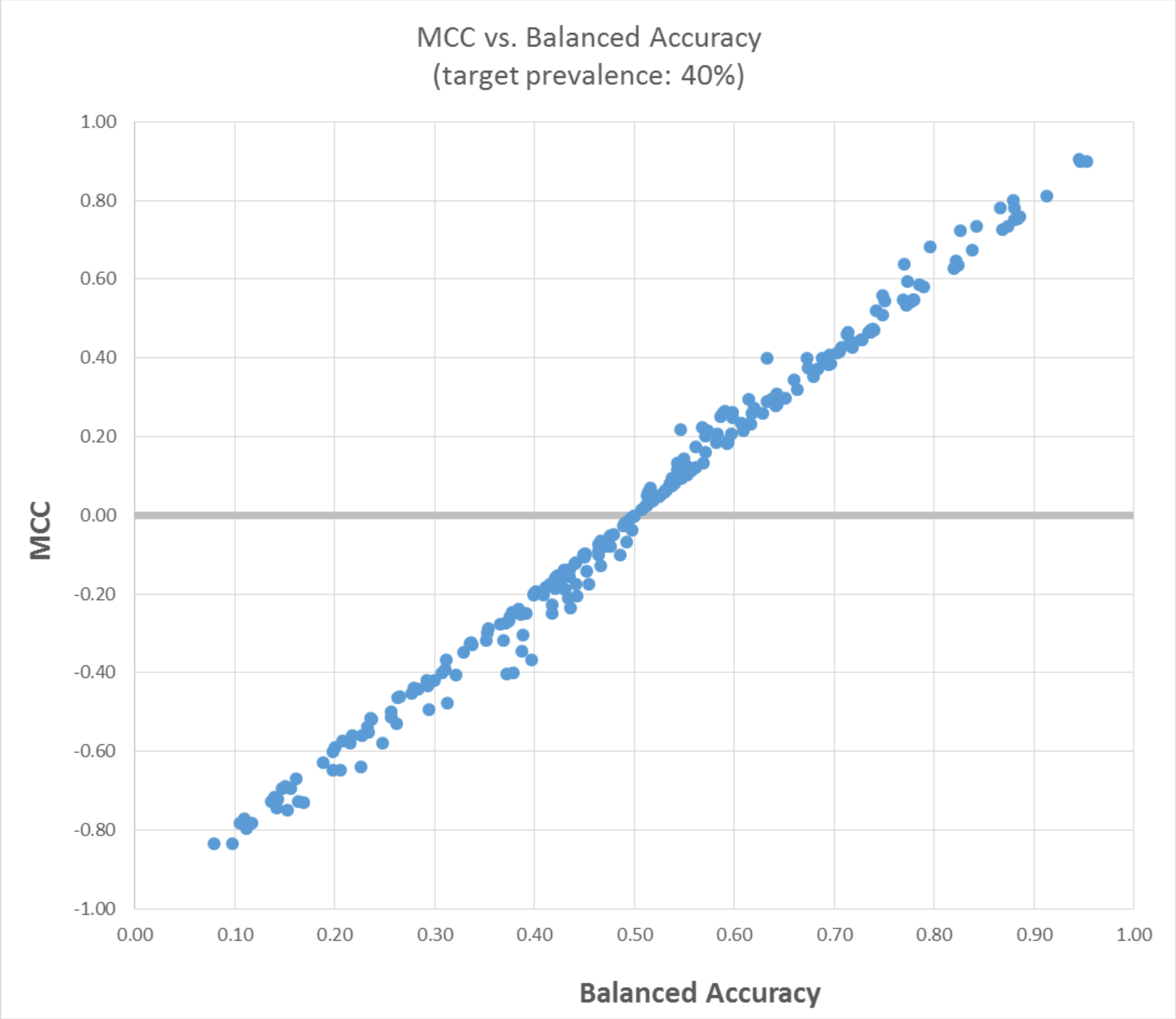
MCC vs. Accuracy





MCC vs. Balanced Accuracy





Log Loss

- The confusion matrix, accuracy, balanced accuracy, F1 score, and MCC rely on specific cutoff point to be specified to evaluate the classifier
- None of these incorporate the predictive probabilities
  - e.g., if cutoff is 0.5, then probabilities of 0.501 and 0.999 are considered equal
- Log loss is a metric that does distinguish the quality of the classifier based on the predicted probabilities
  - If the true case is positive, and the predictive probability is 0.999, this is better then if the predictive probability was 0.501
  - If the true case is negative, and the predictive probability is 0.999, this is worse then if the predictive probability was 0.501

Log loss

$$\text{Log loss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log p_i + (1 - y_i)\log(1 - p_i)\right]$$

- You should recognize the formula from our discussion on deviance residuals and log likelihood for in logistic regression!
- *See example at right.* In these four cases, the log loss is computed assuming the true value is a positive case.
- The goal is to minimize the average of log loss across all predictions.

y_i	Predictive probability	Log loss
1	0.999	0.0004
1	0.501	0.3002
1	0.250	0.6021
1	0.001	3.0000

Cumulative Gains Chart

- Similar to ROC curve, in that it is a graphical technique which evaluates quality of classifier by comparing to a random and perfect
- In contrast to ROC curve (or confusion matrix) which evaluates on the whole data set, cumulative gain charts evaluate model performance in a portion of the data
- Closely related to lift charts (which we will see soon)
- Can be used to support cost benefit analysis in intervention strategies (which we will see soon)
- Note: here it is typical to refer to the data set as a *population*. This is probably due to the fact that cumulative gains charts are often used in evaluating interventions (e.g., marketing campaigns) to people (e.g., customers)

Steps in creating cumulative gains chart

1. Generate predictive probabilities for data set with known outcomes
2. Sort the data set in order from highest to lowest predictive probabilities
3. Group the sorted data into n-tiles (usually deciles); where the first group has the highest probabilities, and the last has the lowest probabilities
4. Compute:
   - percent of target captured by *cumulative* n-tile in the sorted data.
   - percent of target captured by *cumulative* n-tile assuming the model was (1) perfect and assuming (2) it was random
5. Plot the results by n-tile

Example cumulative gains chart

- Assume 10,000 training records have been "scored" (i.e., their predictive probabilities have been calculated based on the model)
- Assume the target prevalence is 24% (2,400 positive cases)
- Assume we are using deciles
- Each decile will have the 1,000 records
  - Decile 1 will have the 1,000 records with the highest probabilities
  - Decile 10 will have the 1,000 records with the lowest probabilities

grouped into deciles

Decile	Predicted probability
1	0.999
1	0.987
1	0.986
1	0.981
…	….
2	0.785
2	0.769
2	0.742
2	0.741
…	…

sorted

Compute the *per decile* statistics

- Determine the quantity of target (positive outcomes) in each decile based on your sorting and grouping into deciles
- Determine what the expected quantity would be if you didn't use a model and only grouped into deciles randomly.
- Determine what the value would have been if your model was perfect ("wizard")

$$Yp = 0.3789 - 0.1621\, x_1 - 0.6200\, x_2$$
$$0.0414\, x_4 - 0.0567 x_5 + 0.8012\, x_6 +$$
$$0.5667\, x_1 x_2 + 0.0277 x_1 x_3 - 0.0953\, x_1 x_4$$

Decile	Quantity of population	Quantity of target captured (model)	Quantity of target captured (random)	Quantity of target captured (wizard)
1	1000	792	240	1000
2	1000	528	240	1000
3	1000	432	240	400
4	1000	216	240	0
5	1000	204	240	0
6	1000	144	240	0
7	1000	72	240	0
8	1000	6	240	0
9	1000	4	240	0
10	1000	2	240	0

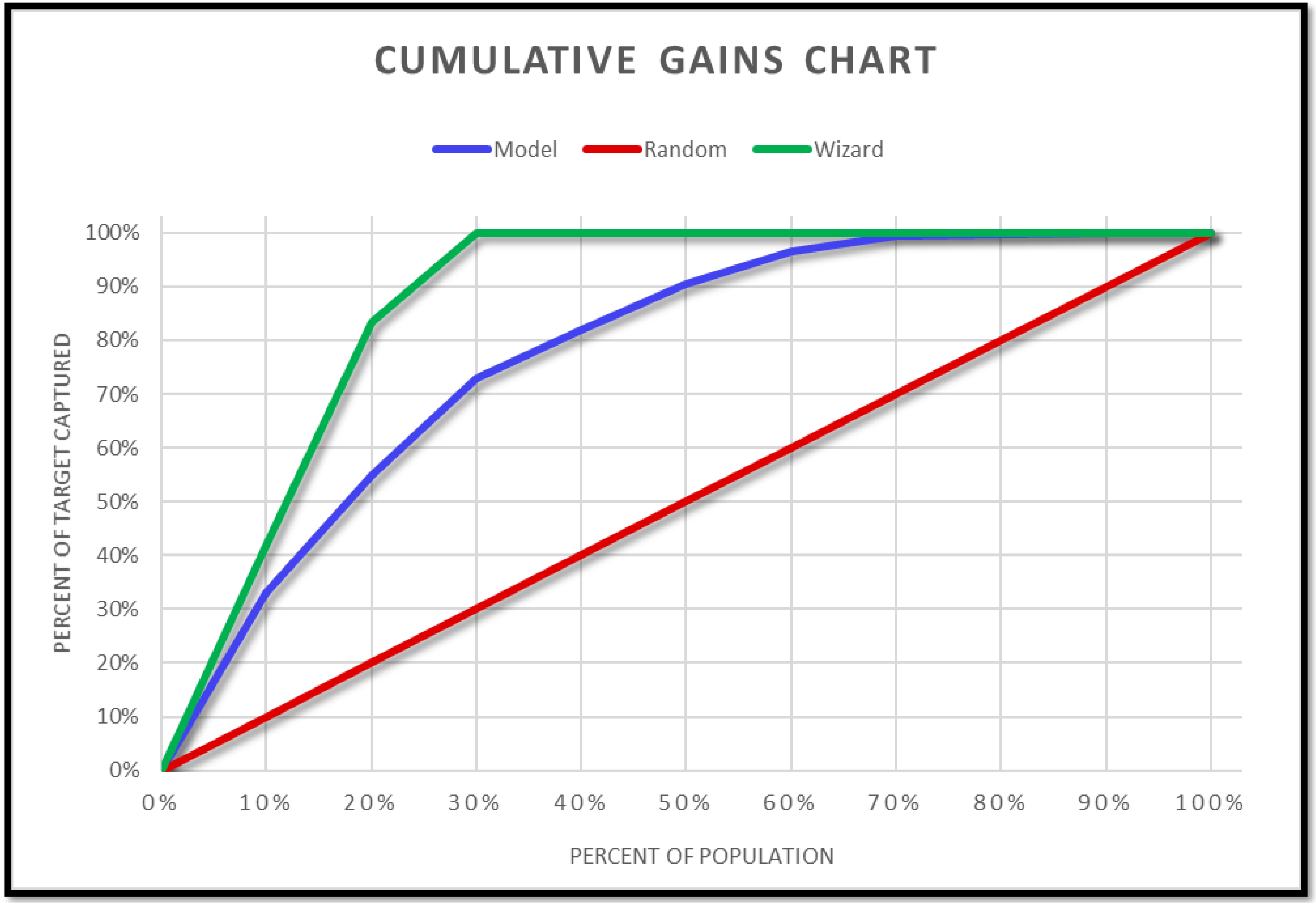
Example cumulative gains chart

- Compute the statistics associated with the *cumulative* deciles
- Quantity of target captured is the sum of positive cases for cumulative deciles
- Do this for the random and wizard decile statistics too

Cumulative Decile	Quantity	Percent of population	Quantity of target captured (model)	Percent of target captured (model)
1	1000	10%	792	33%
2	2000	20%	1320	55%
3	3000	30%	1752	73%
4	4000	40%	1968	82%
5	5000	50%	2172	91%
6	6000	60%	2316	97%
7	7000	70%	2388	100%
8	8000	80%	2394	100%
9	9000	90%	2398	100%
10	10000	100%	2400	100%

Cumulative Gains Chart

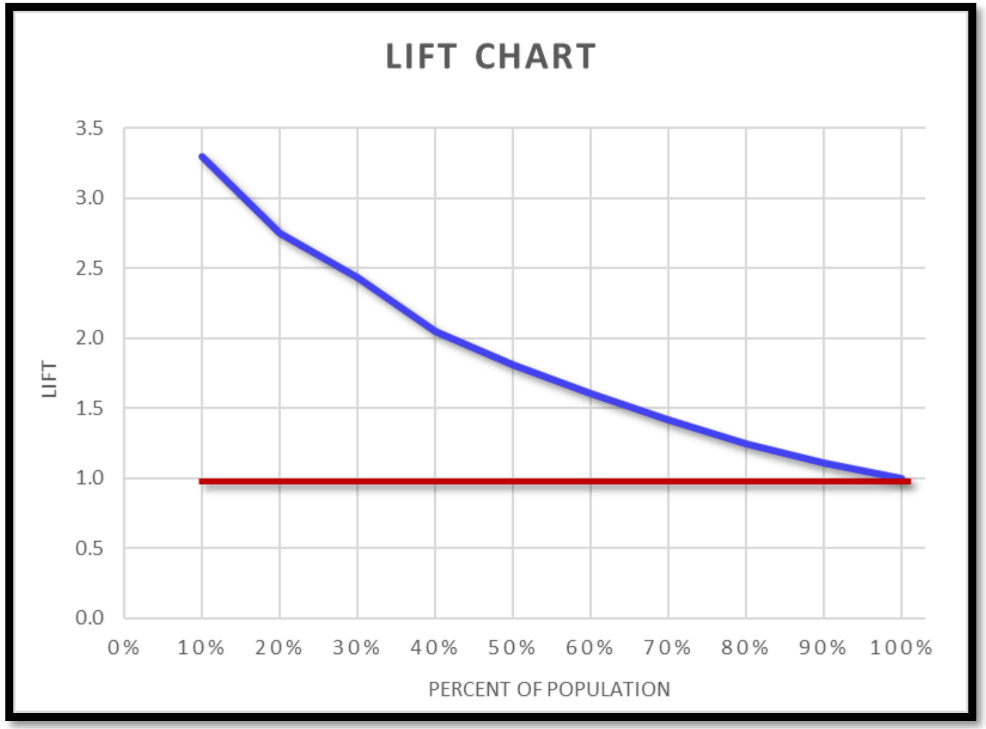
- Plot the percentages captured vs. the percent of the population
- The blue curve can never be better than the green curve
- The blue curve "hopefully" is better than the red curve, but it could be worse!



Lift chart

- Very simple extension based on the statistics computed during the cumulative gains chart process
- Plot of ratio: *cumulative percent target captured using the model* to *cumulative percent target captured using the model* by n-tile

Cumulative Decile	Random	Model	Lift
1	10%	33%	3.30
2	20%	55%	2.75
3	30%	73%	2.43
4	40%	82%	2.05
5	50%	91%	1.81
6	60%	97%	1.61
7	70%	100%	1.42
8	80%	100%	1.25
9	90%	100%	1.11
10	100%	100%	1.00



Additional remark

- The information from the cumulative gains chart can help a company decide on how to use the model to help with an intervention strategy
- Given: marginal cost associated with intervention (e.g., cost to contact a customer via mail); expected $ benefits or losses with intervention of correct target
- Then: it is straight-forward to project expected profits by cumulative n-tile.