Presenter Introduction

Name: Philip Sierpinski

Background: Bachelor of Mathematics

Position: Solution Advisor for Advanced Analytics & AI

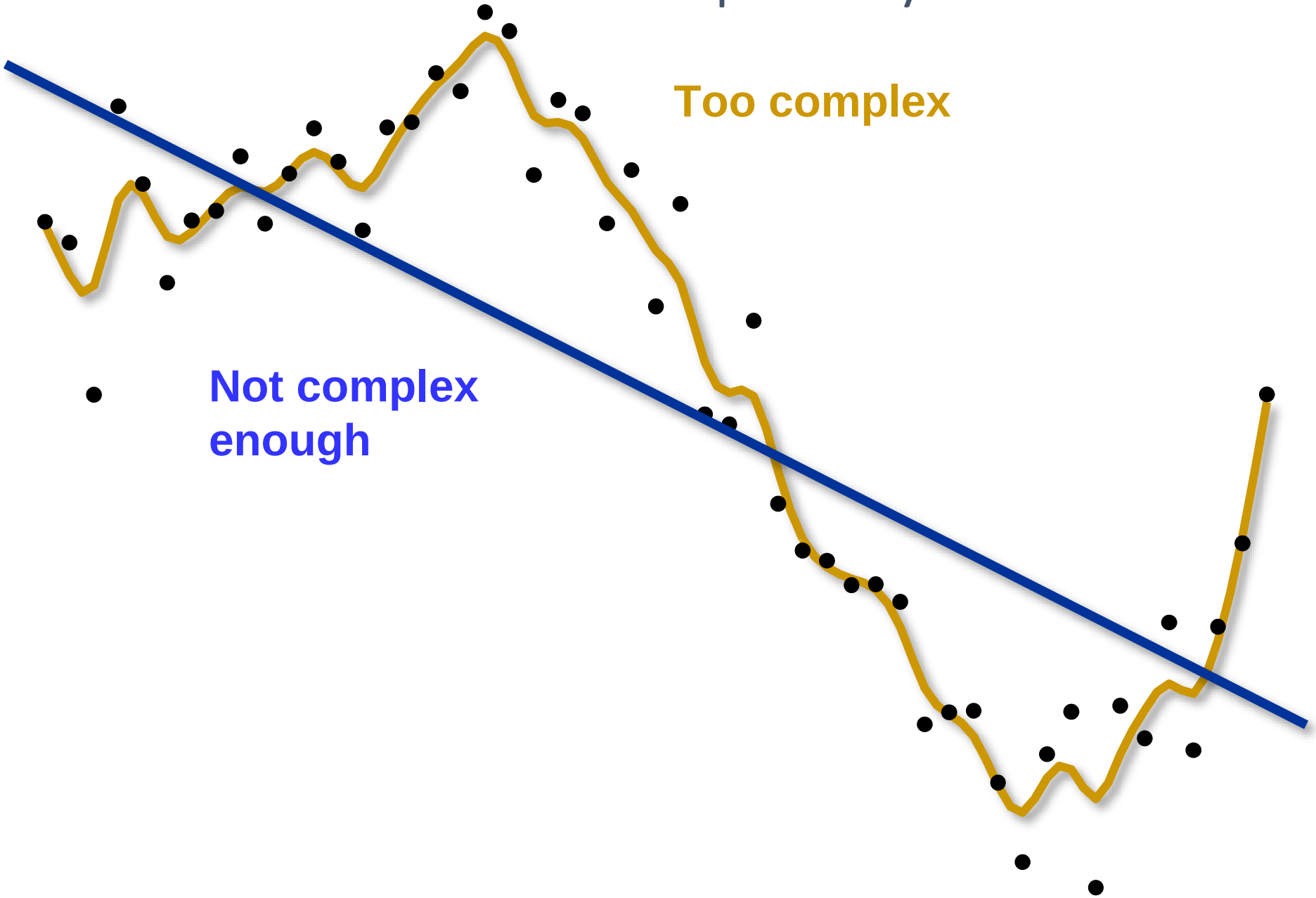
Philip.Sierpinski@sas.com

Fit Statistics

Validating your models is a crucial step of the modelling process.

With a lot of different fit statistics available to us – How do we choose which one to use? How do they differ? Are there any pitfalls we should be aware of?

Model Complexity



...

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	**Accuracy/Misclassification** **Profit/Loss** **Inverse prior threshold**
Rankings	ROC Index (concordance) Gini coefficient
Estimates	Average squared error SBC/Likelihood

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
Rankings	**ROC Index (concordance)** **Gini coefficient**
Estimates	Average squared error SBC/Likelihood

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
Rankings	ROC Index (concordance) Gini coefficient
Estimates	**Average squared error** **SBC/Likelihood**

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	**Accuracy/Misclassification** **Profit/Loss** **Inverse prior threshold**
Rankings	ROC Index (concordance) Gini coefficient
Estimates	Average squared error SBC/Likelihood

Misclassification

Misclassifying an observation means that you have incorrectly predicted the outcome for that observation.

Used for data that aims to predict an event occurring or not.

A smaller misclassification rate is thus better.

Example: Image recognition

We have images of cats and dogs and have trained a convolutional neural network to classify these images.

We would like to validate how well our model is performing by looking at the misclassification rate.

Example: Image Recognition

Actual: [Cat, Cat, Cat, Cat, Cat, Cat, Cat, Cat, Dog, Dog, Dog, Dog]

Predicted: [Dog, Dog, Cat, Cat, Cat, Cat, Cat, Cat, Dog, Dog, Dog, Cat]

Example: Image Recognition

	Predicted Cat	Predicted Dog
Actual Cat	6	2
Actual Dog	1	3

	Predicted P	Predicted N
Actual P	TP	FN
Actual N	FP	TN

Prediction Cut-offs

Allows you to change the distribution of the TP,FN,FP,TN. Can be utilized if you are only interested in detecting for example positives (maybe for virus tests)

Changes in the probability cut-off value (numeric value between 0 and 1) decides if a prediction should be counted as an event (yes, infected) or not (no, not infected)

Pitfalls of the misclassification rate

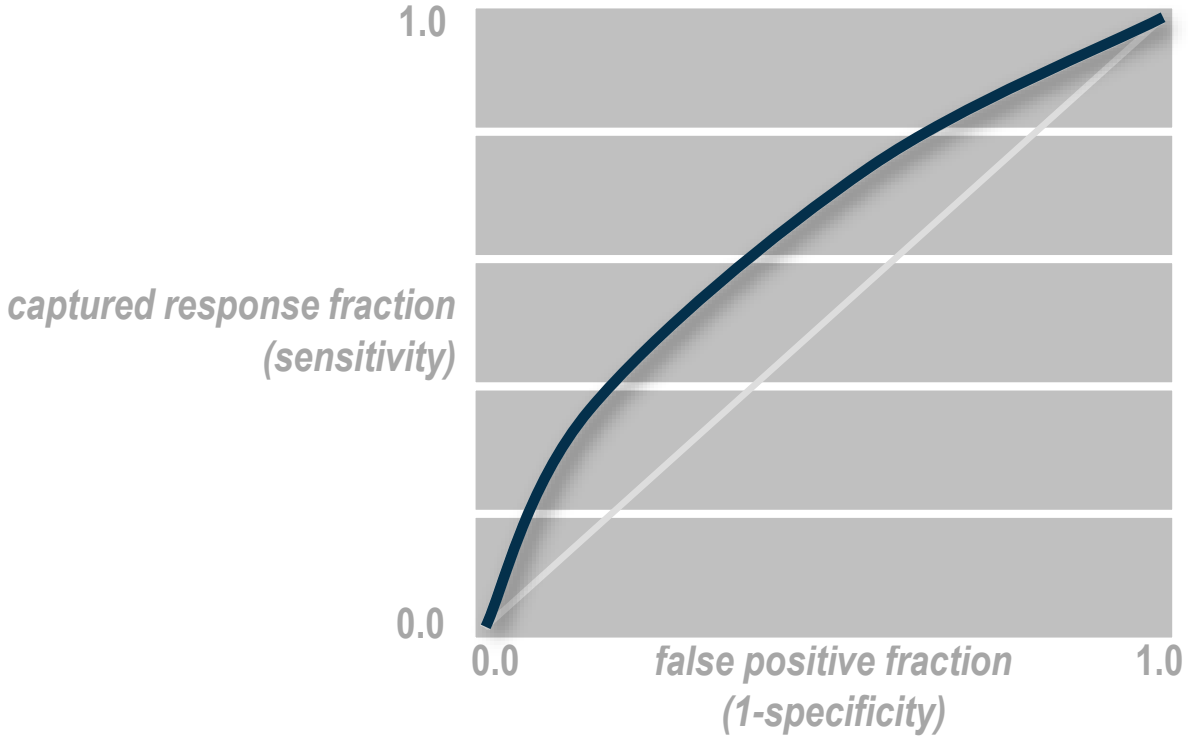
Can in some cases lead to misleading results

Example: Unbalanced data sets

Summary Statistics Summary

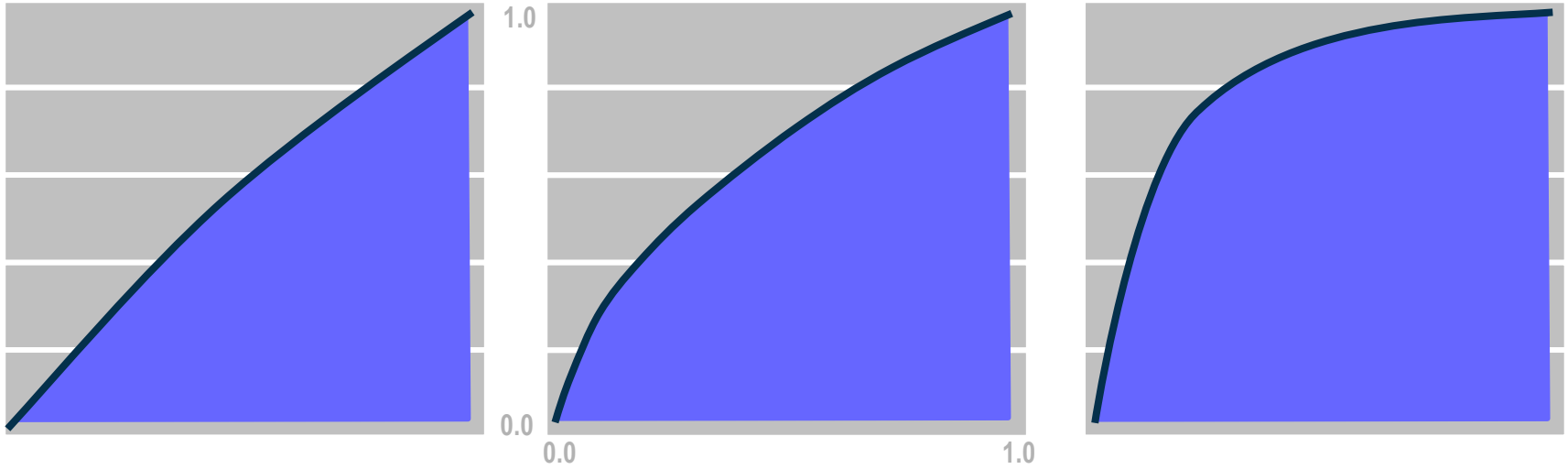
Prediction Type	*Statistic*
Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
Rankings	**ROC Index (concordance)** **Gini coefficient**
Estimates	Average squared error SBC/Likelihood

ROC Chart



The ROC chart illustrates a tradeoff between a captured response fraction and a false positive fraction.

Statistical Graphics: ROC Index



weak model
ROC Index < 0.6

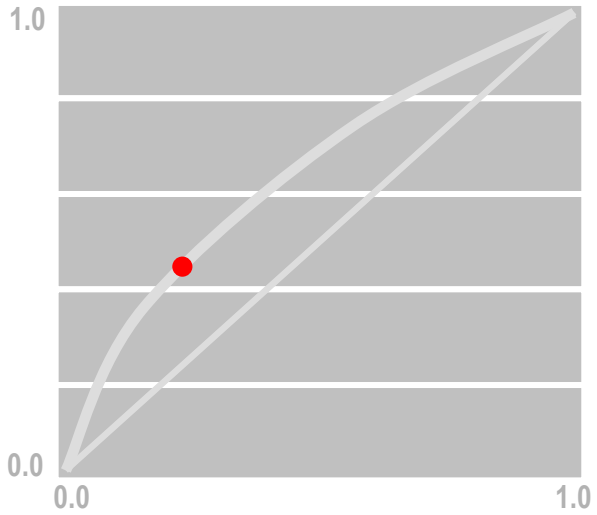
strong model
ROC Index > 0.7

ROC

$$\text{Sensitivity(y-axis)} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

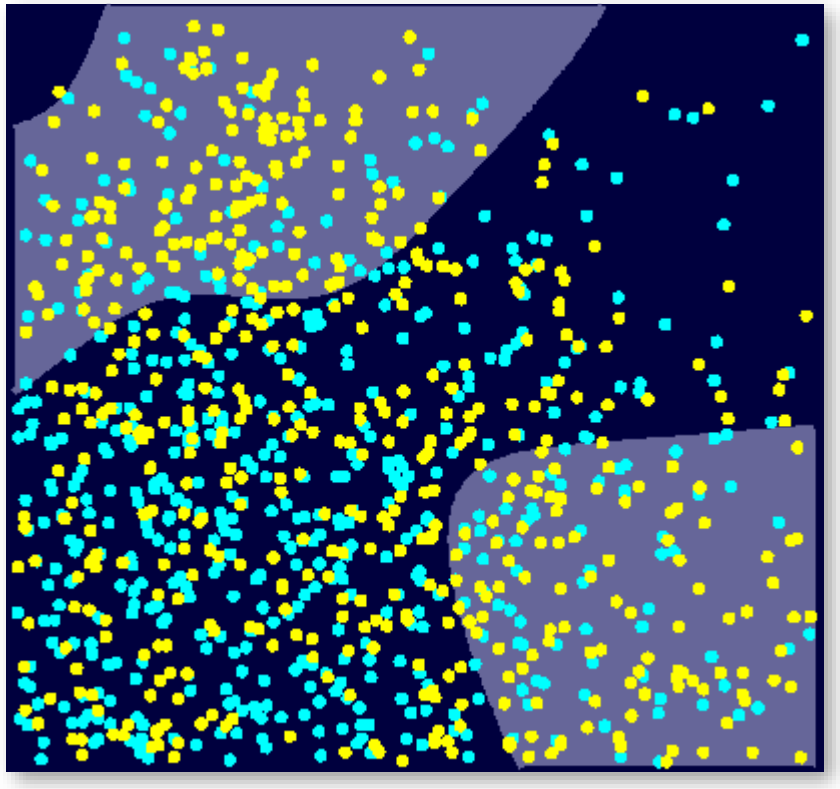
$$\text{(1-Specificity)(x-axis)} = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

Statistical Graphics: ROC Chart

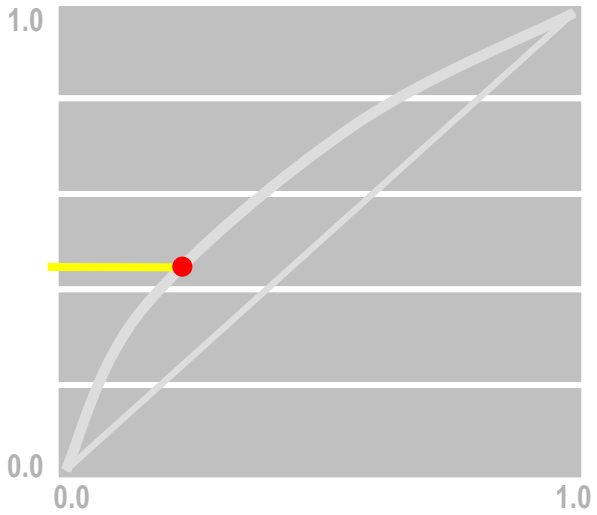


For example, this point on the ROC chart corresponds to the 40% of cases with the highest predicted values.

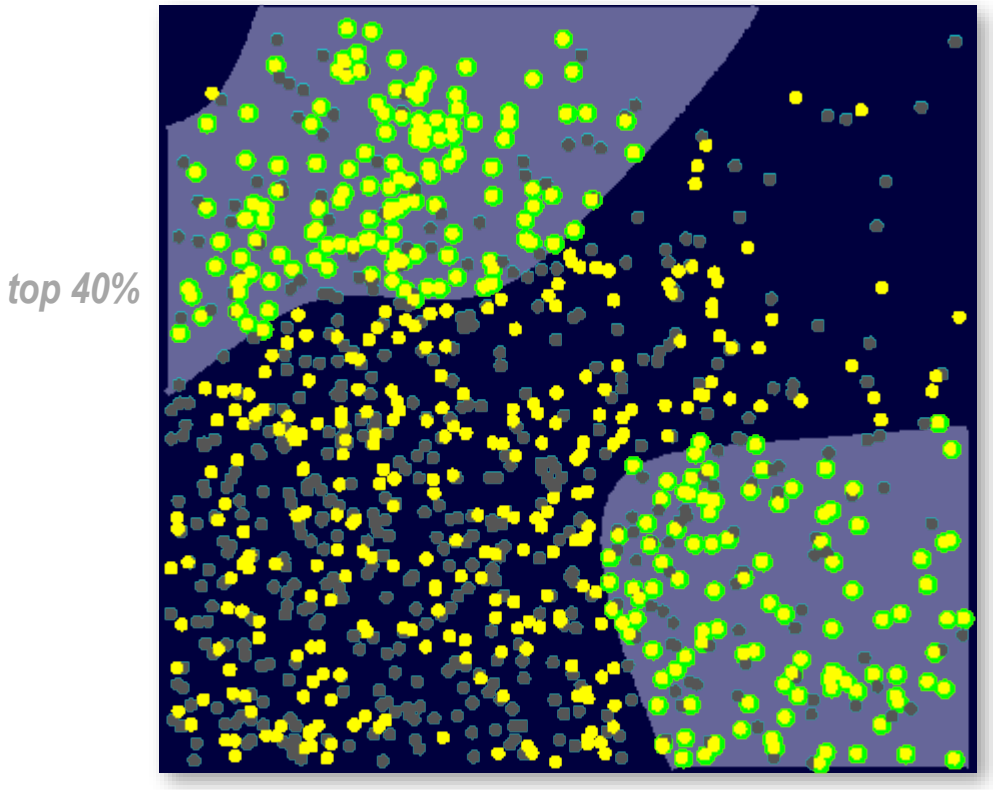
top 40%



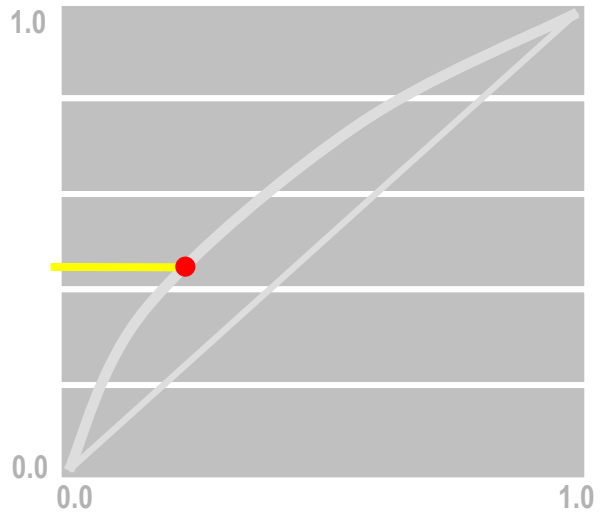
Statistical Graphics: ROC Chart



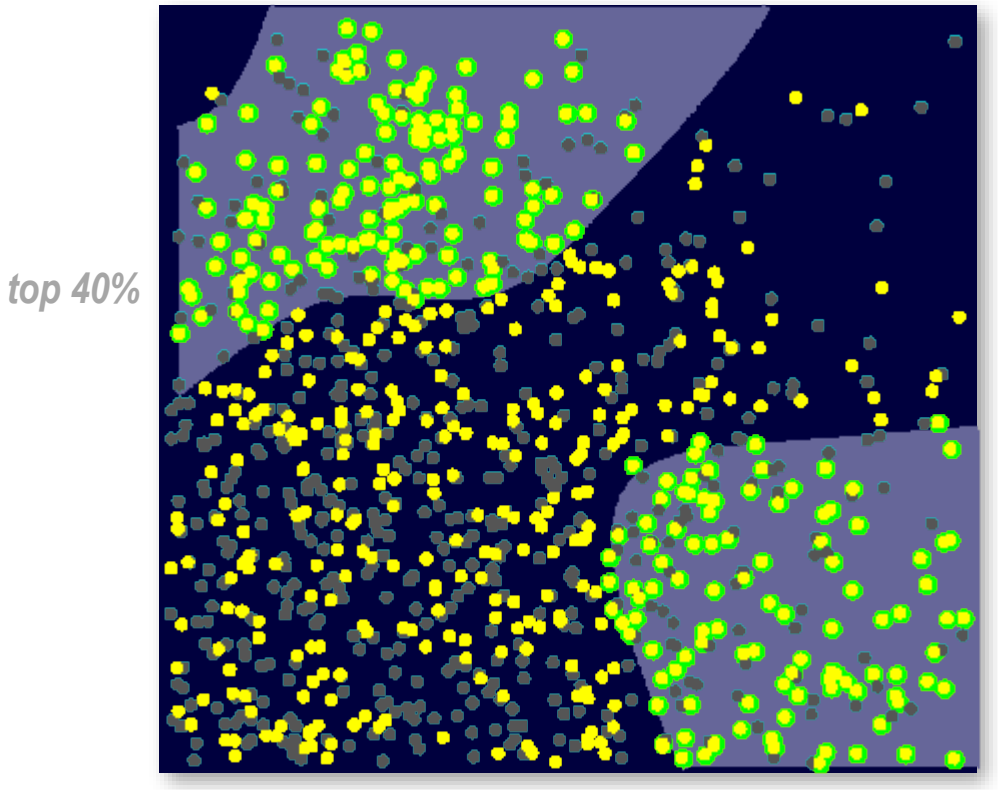
The *y*-coordinate shows the fraction of primary outcome cases captured in the top 40% of all cases.



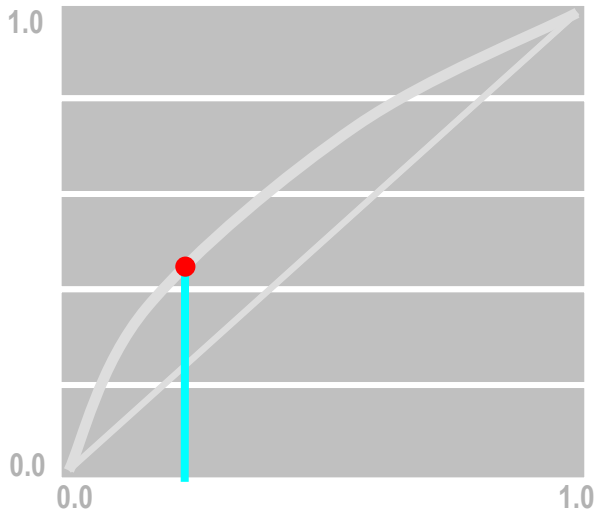
Statistical Graphics: ROC Chart



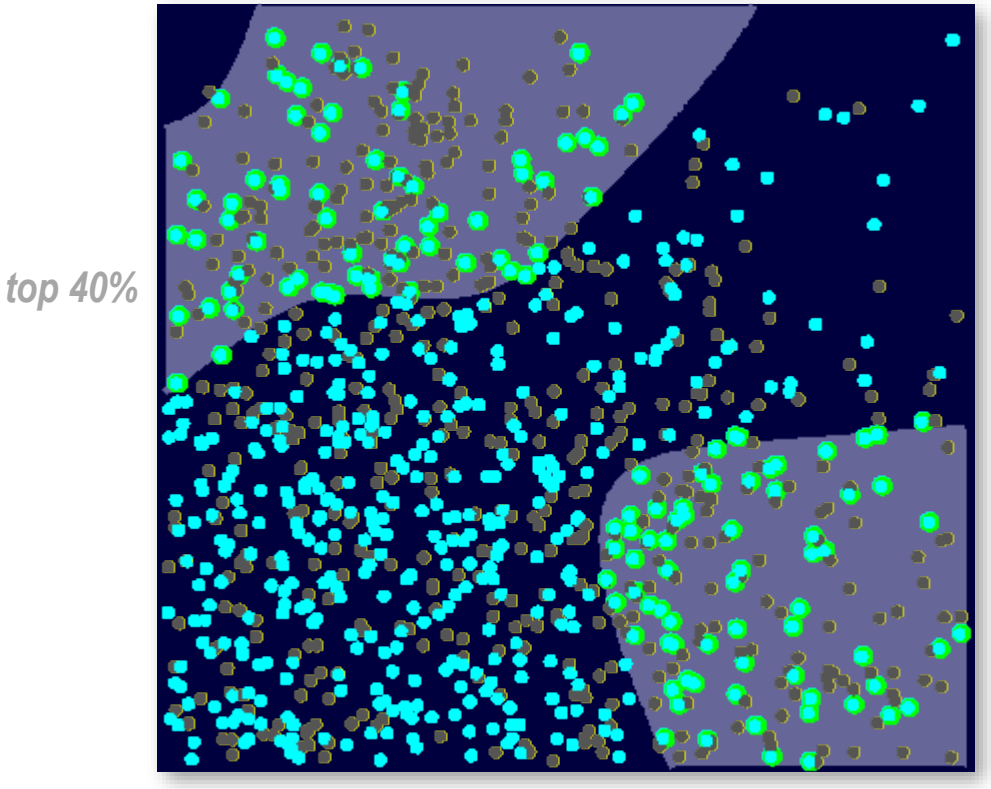
The *y*-coordinate shows the fraction of *primary* outcome cases captured in the top 40% of all cases.



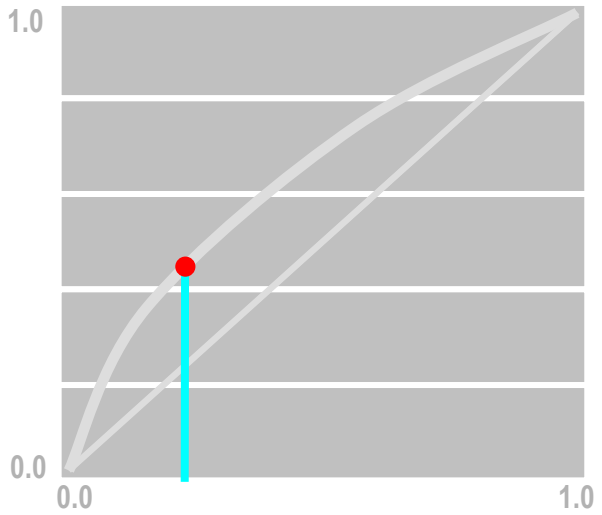
Statistical Graphics: ROC Chart



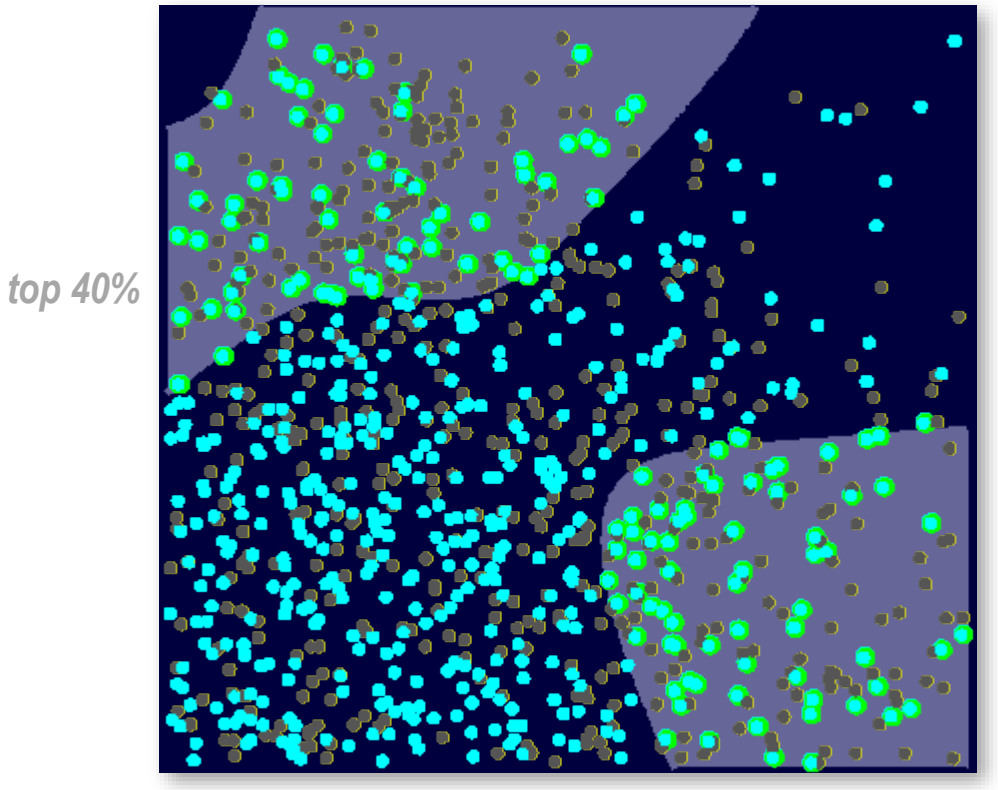
The *x*-coordinate shows the fraction of *secondary* outcome cases captured in the top 40% of all cases.



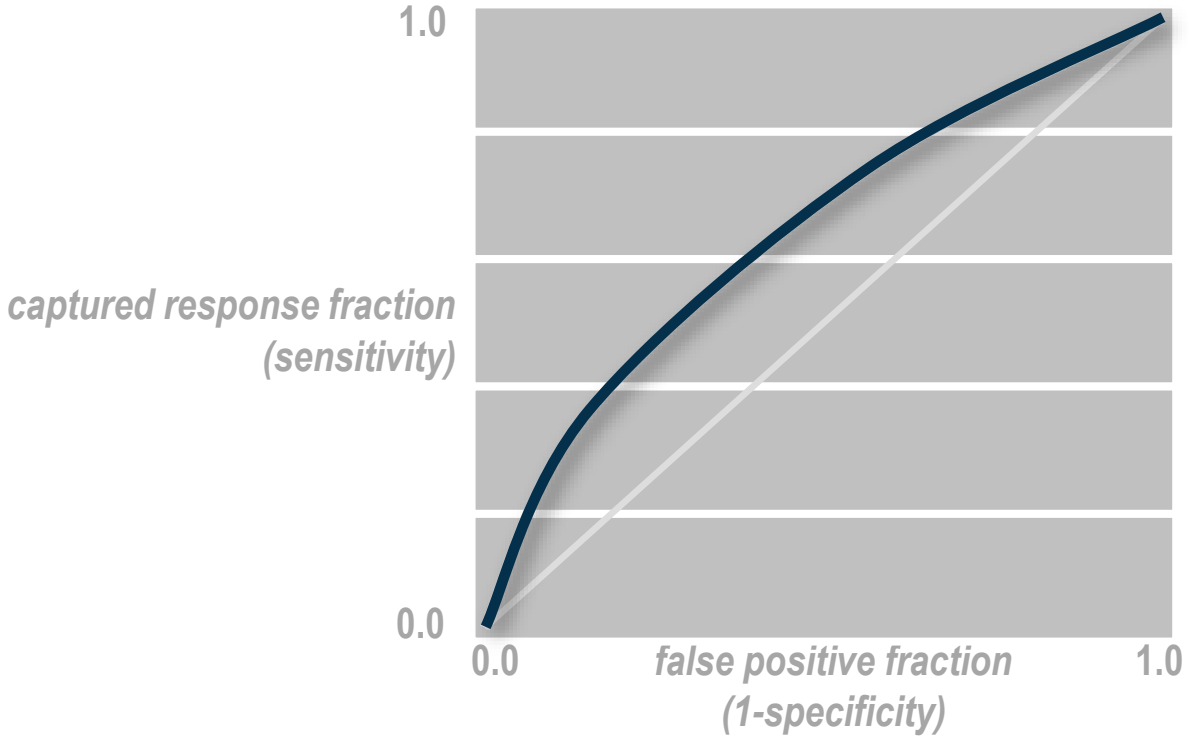
Statistical Graphics: ROC Chart



The *x*-coordinate shows the fraction of *secondary* outcome cases captured in the top 40% of all cases.



ROC Chart



The ROC chart illustrates a tradeoff between a captured response fraction and a false positive fraction.

Pitfalls of the ROC curve

- Any attempt to summarize the ROC curve into a single number loses information about the pattern of tradeoffs of the particular discriminator algorithm
- AUC estimates are quite noisy
- Sometimes it can be more useful to look at a specific region of the ROC Curve rather than at the whole curve

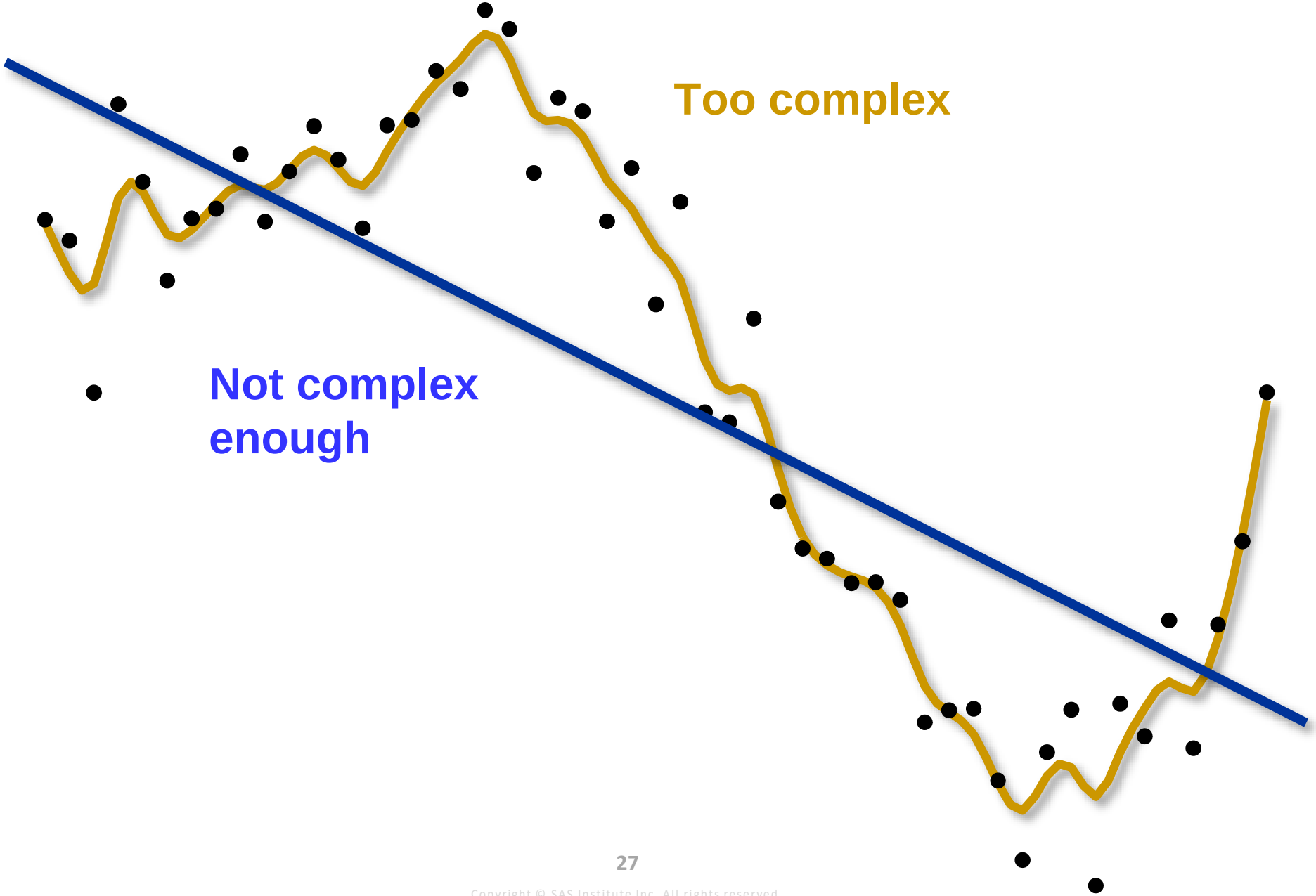
Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
Rankings	ROC Index (concordance) Gini coefficient
Estimates	**Average squared error** **SBC/Likelihood**

Average Square Error

- Main usage is for estimate predictions.
- Often used for regression analysis.

Model Complexity



Pitfalls of ASE

- Outliers heavily influence the statistic

Summary

Three different prediction types: Decisions, Rankings and Estimates.

Depending on what the goal of the model is – use a fit statistic that is favorable for that case.

Be aware of certain pitfalls that apply to the chosen statistic.