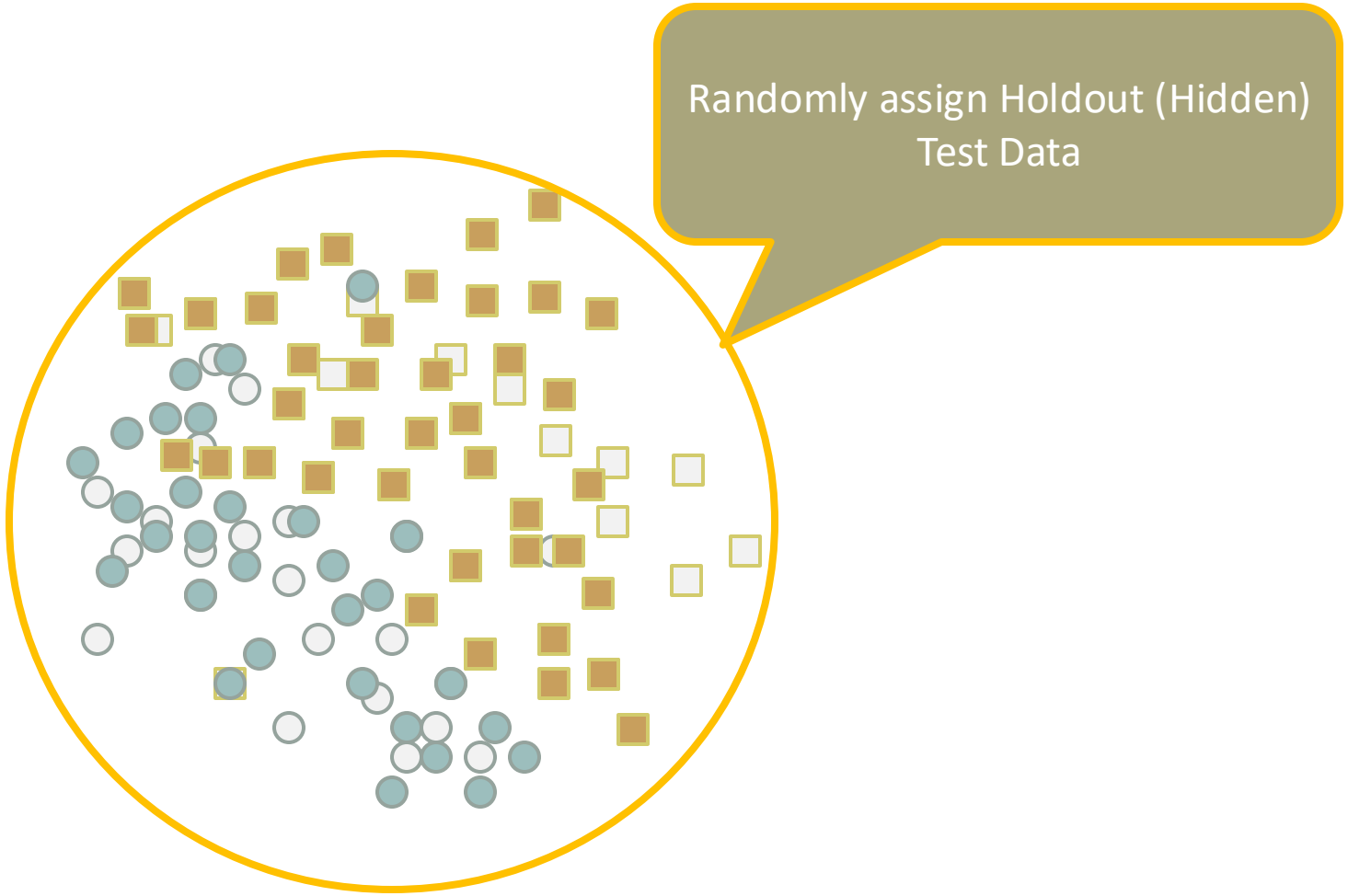
Over-fitting and Confusion Matrix

Evaluate Model

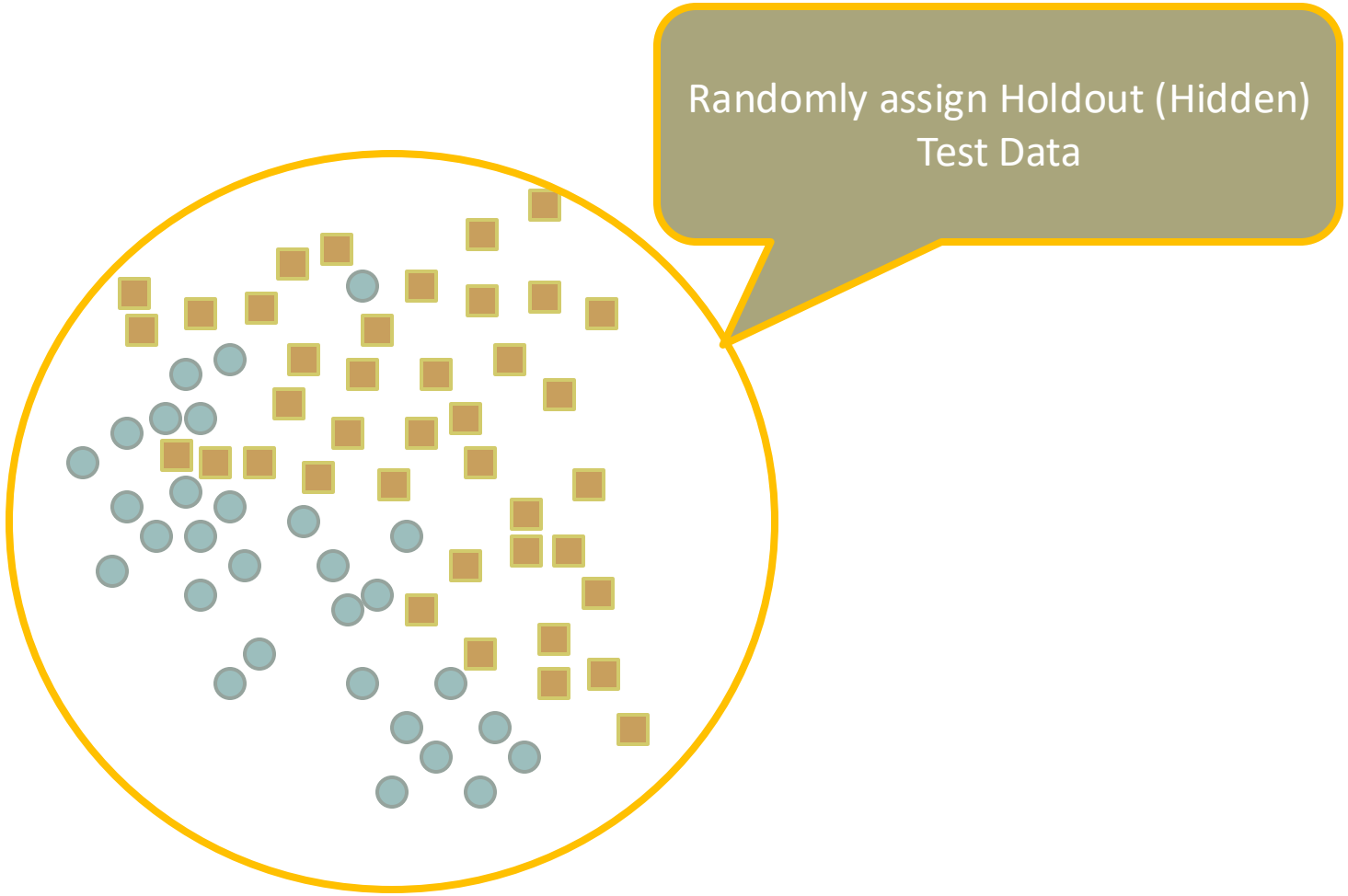
- The following segment will use an over-fitting example to explain the following concepts:
  - Modeling Data
    - Training Data
    - Test Data
  - Model (Hypothesis)
  - Over-fitting
  - Model Accuracy
  - Confusion Matrix (Classification Matrix)
    - True Positive
    - False Positive
    - True Negative
    - False Negative

Evaluate Model: All Data

Evaluate Model: Test Data

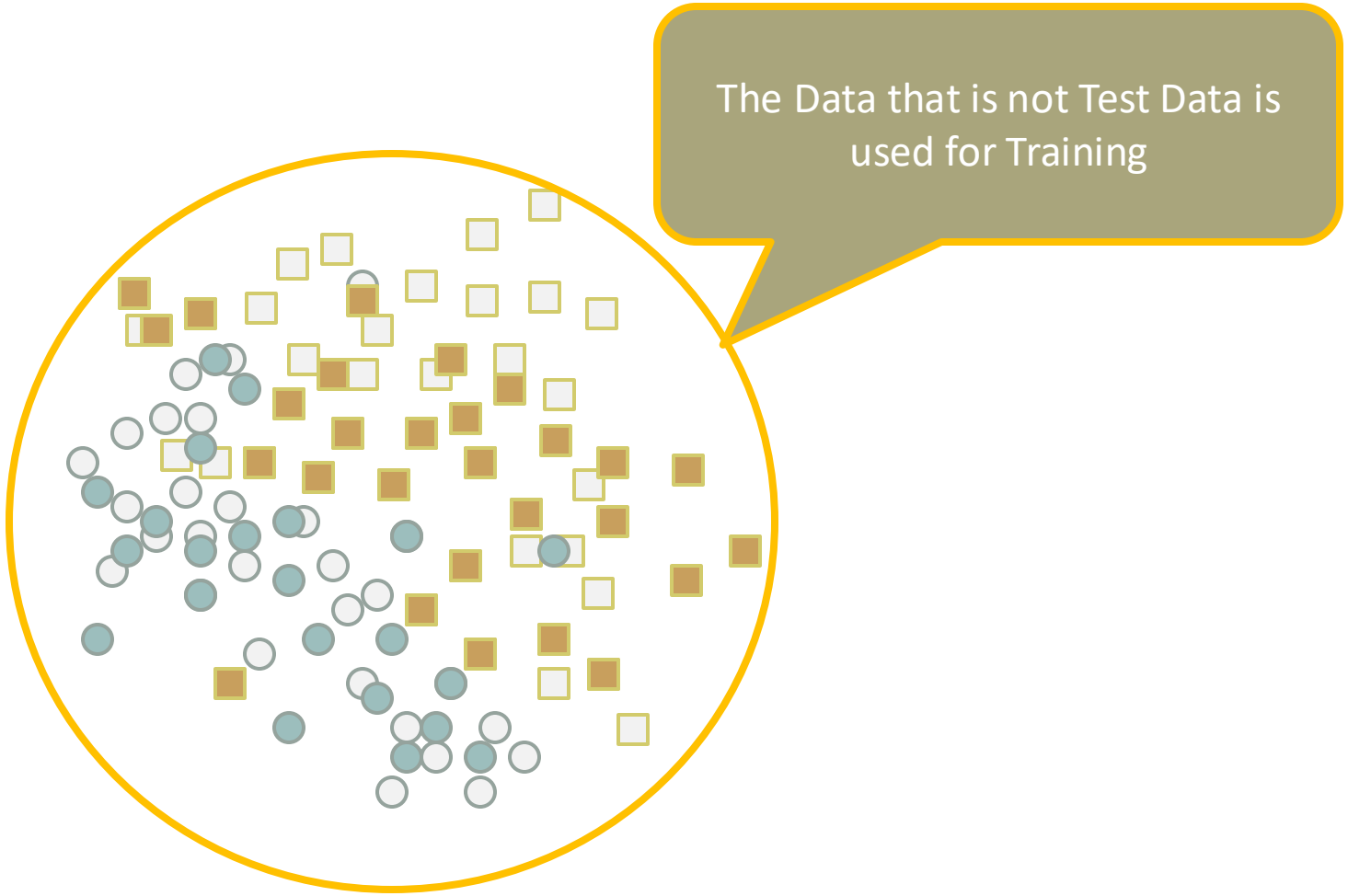


Evaluate Model: Test Data

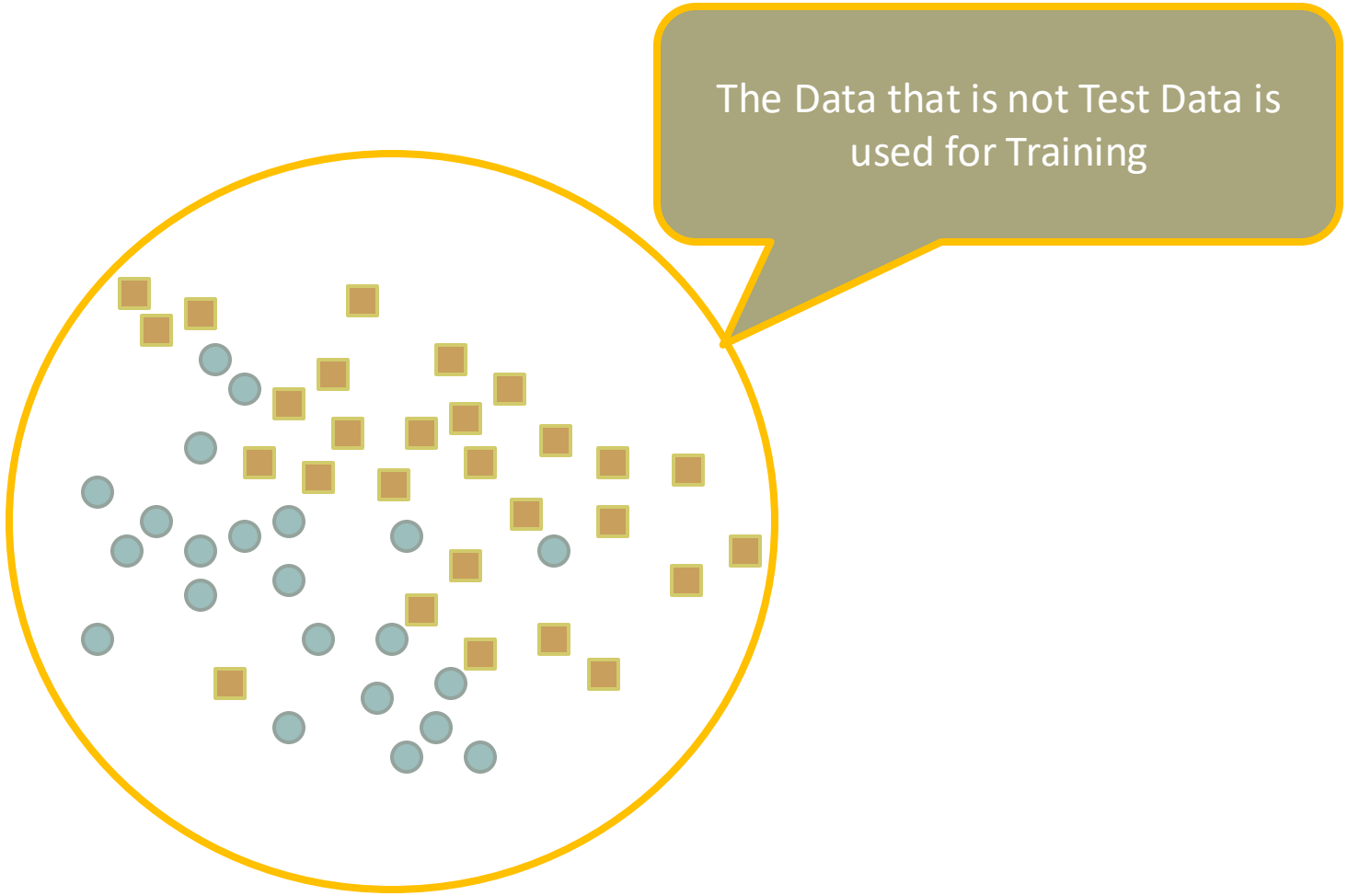


Evaluate Model: All Data

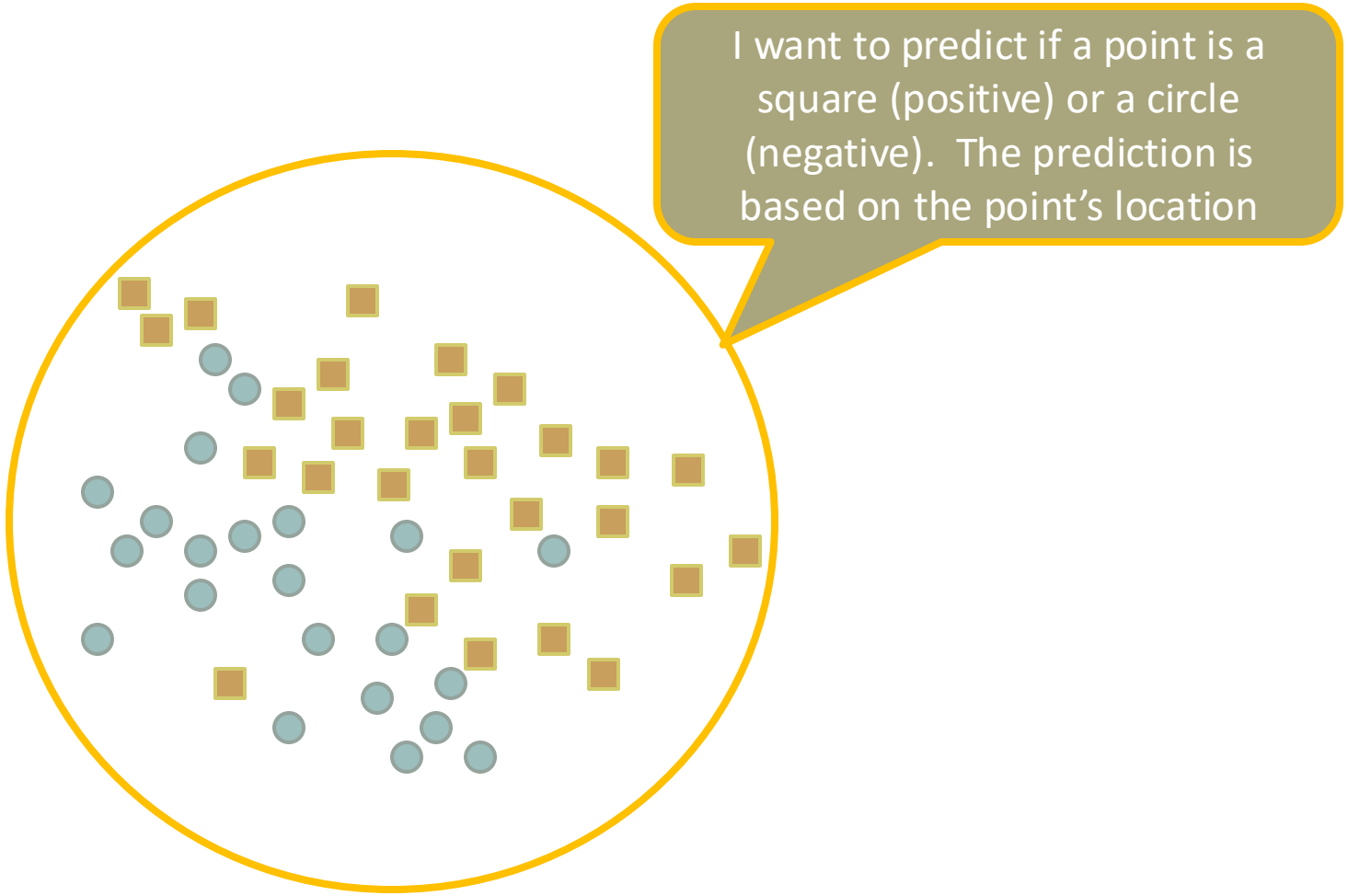
Evaluate Model: Training Data



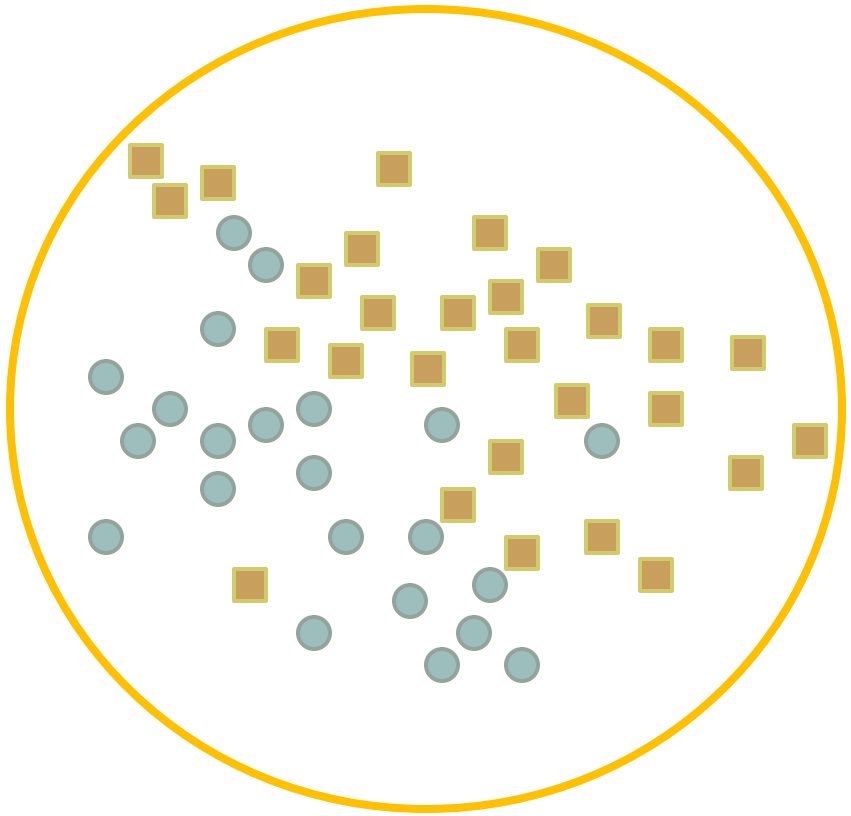
Evaluate Model: Training Data



Evaluate Model: Training



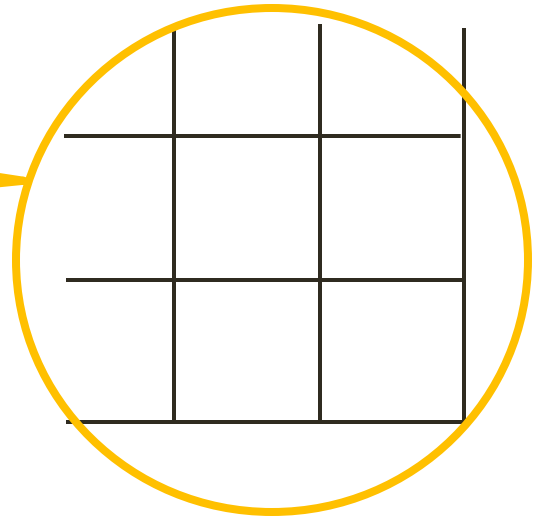
Evaluate Model: Training

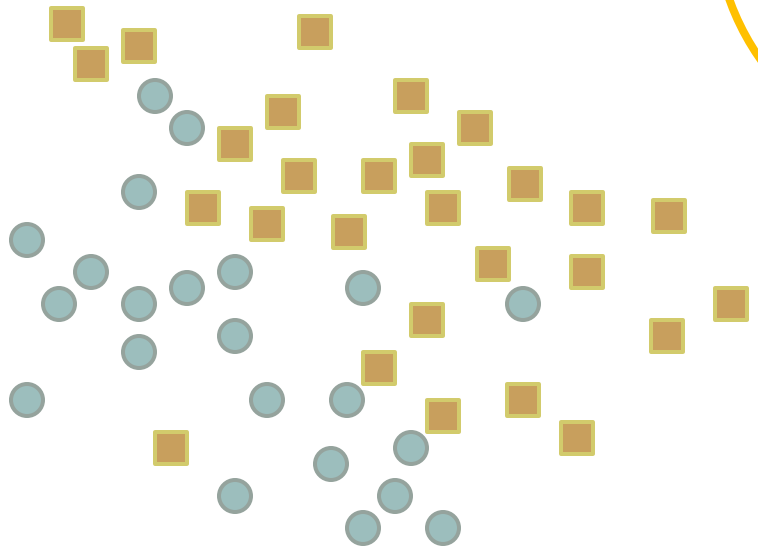


isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

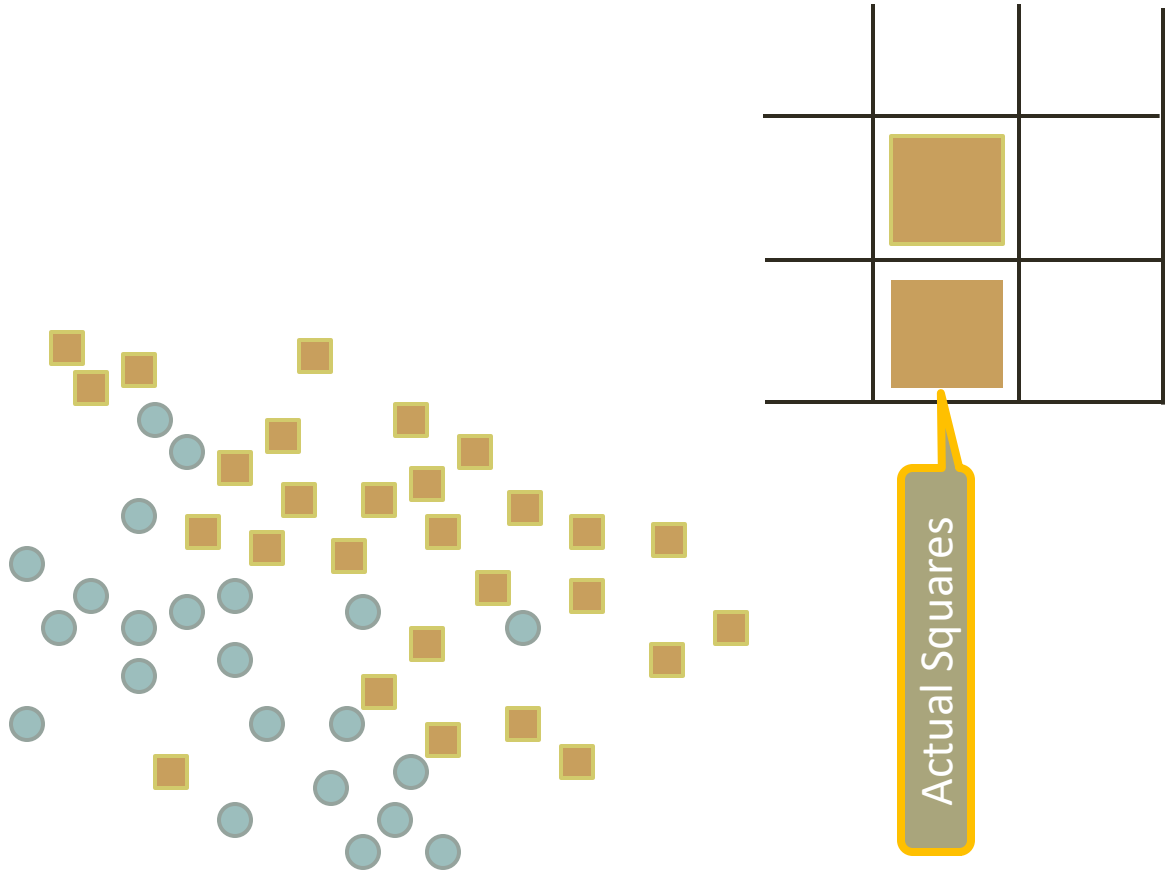
Confusion Matrix (Classification Matrix):
Compare Squares and Circles with
Predicted Squares and Circles





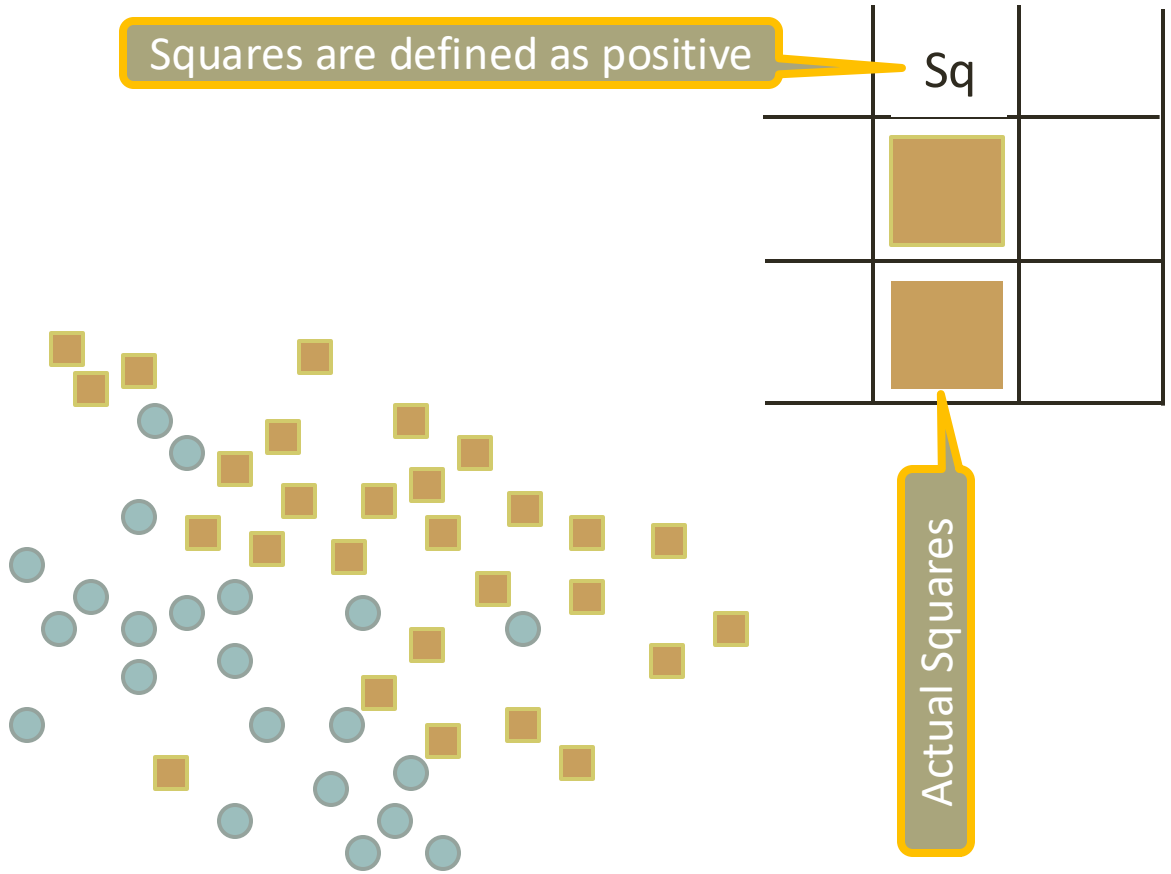
isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix



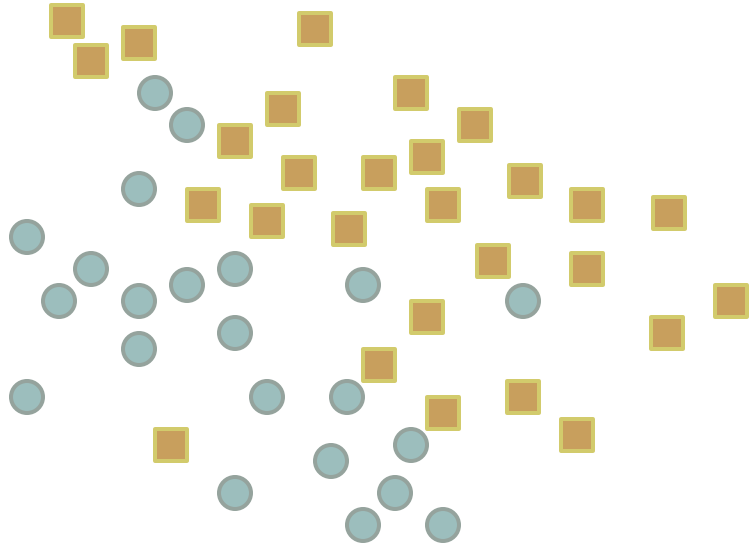
isSquare ~ xLocation + yLocation

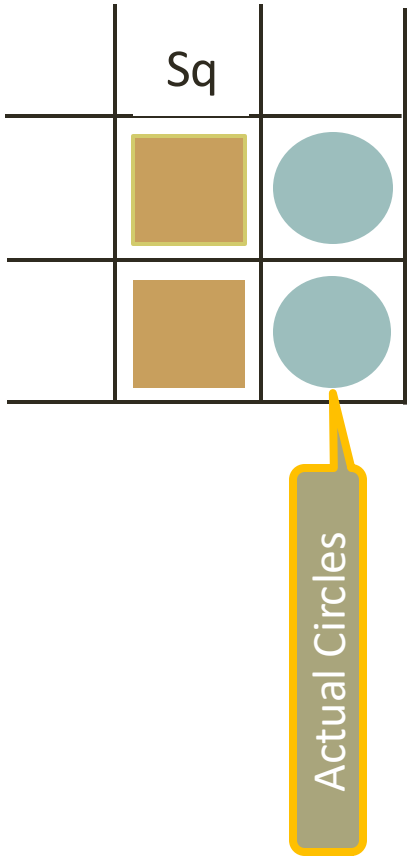
Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

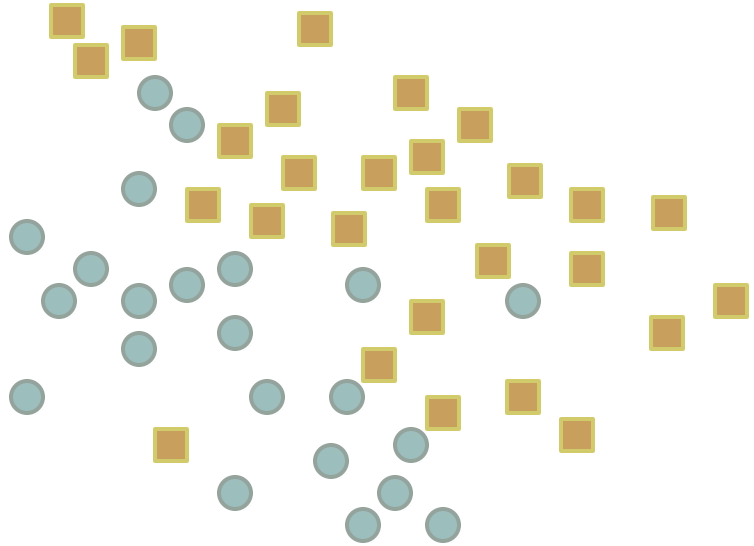
Evaluate Model: Confusion Matrix

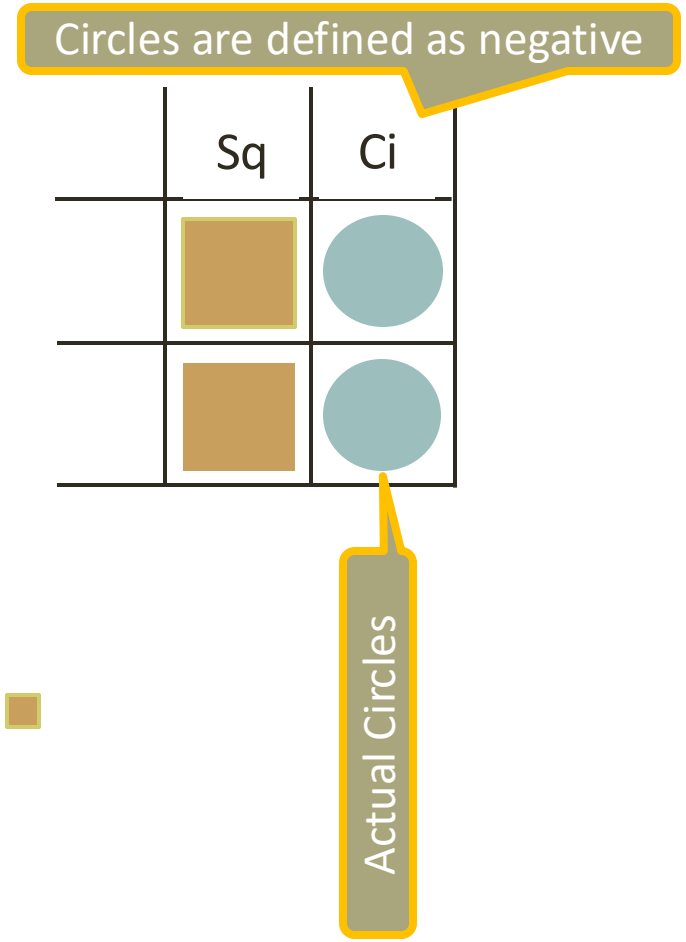




isSquare ~ xLocation + yLocation

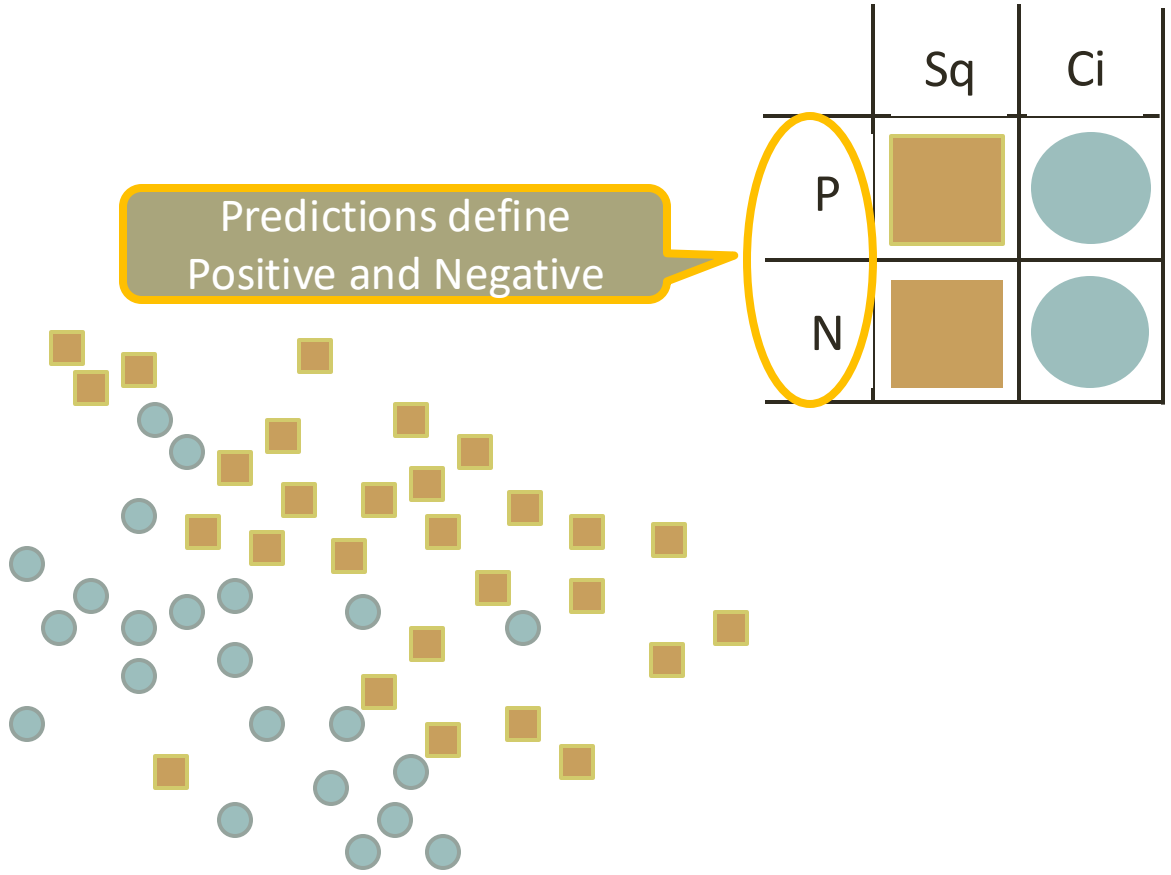
Evaluate Model: Confusion Matrix





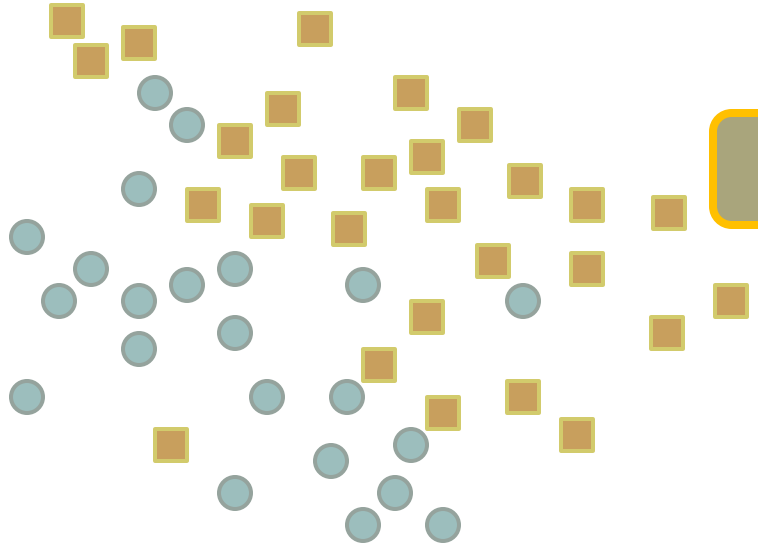
isSquare ~ xLocation + yLocation

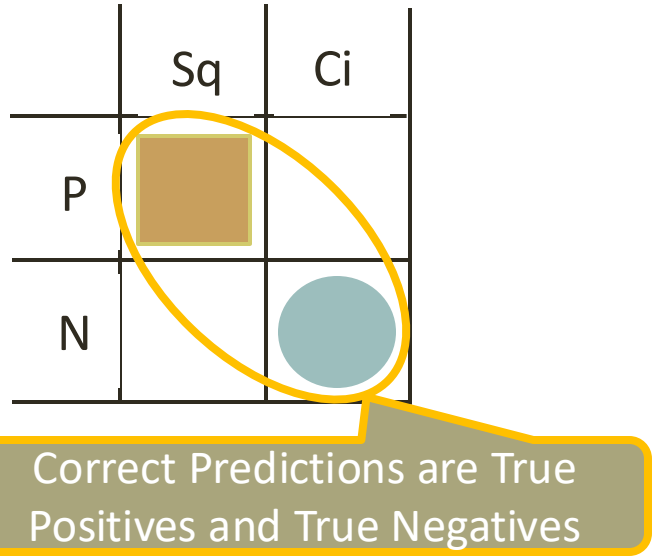
Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

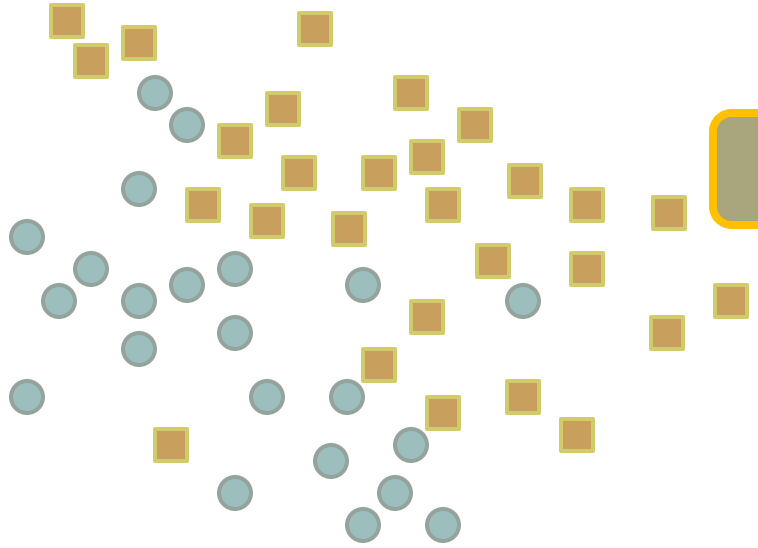
Evaluate Model: Confusion Matrix

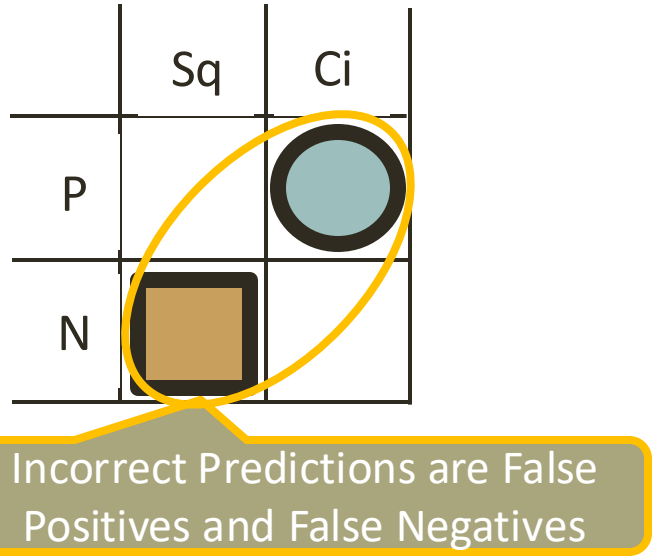




isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

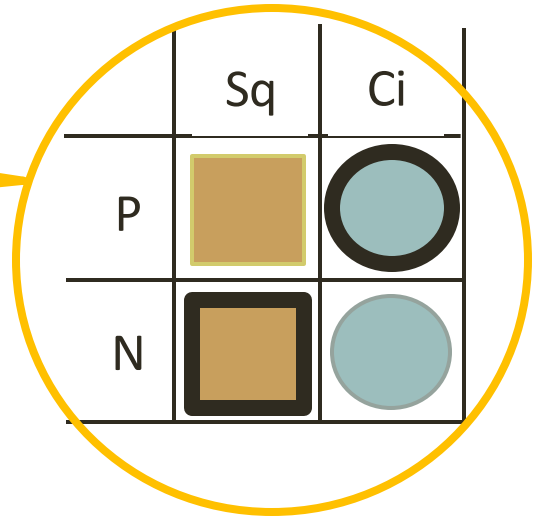


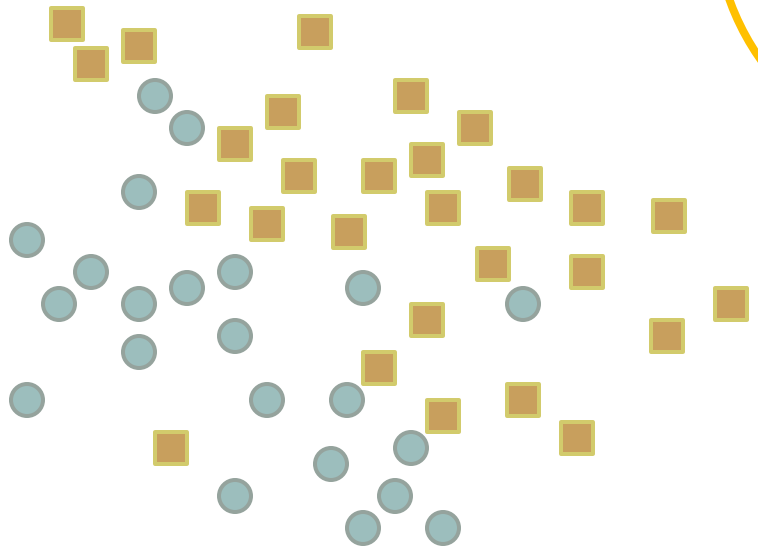


isSquare ~ xLocation + yLocation

Evaluate Model: Confusion Matrix

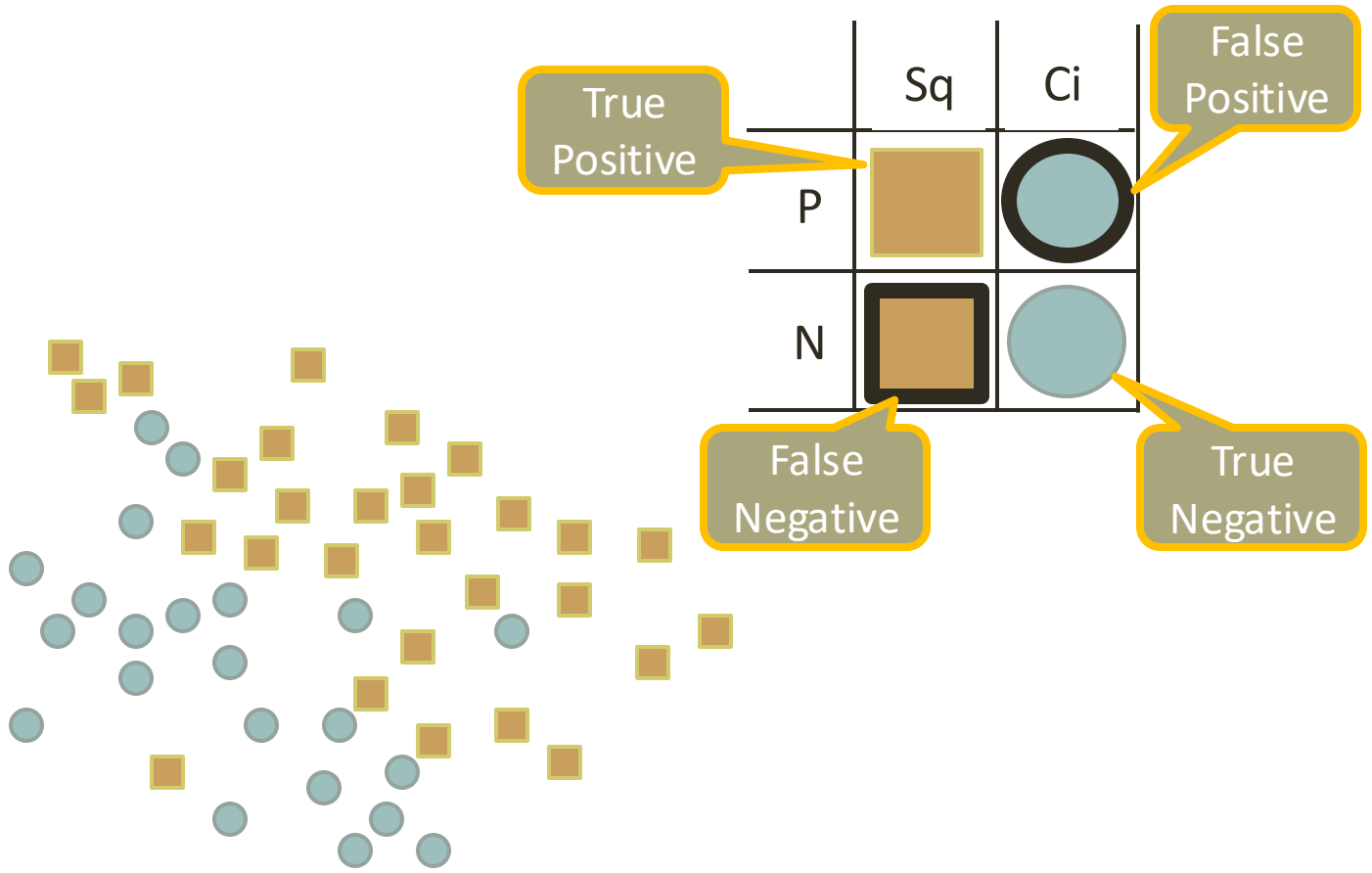
Confusion Matrix (Classification Matrix):
Vertical are actual classes
Horizontal are predicted classes





isSquare ~ xLocation + yLocation

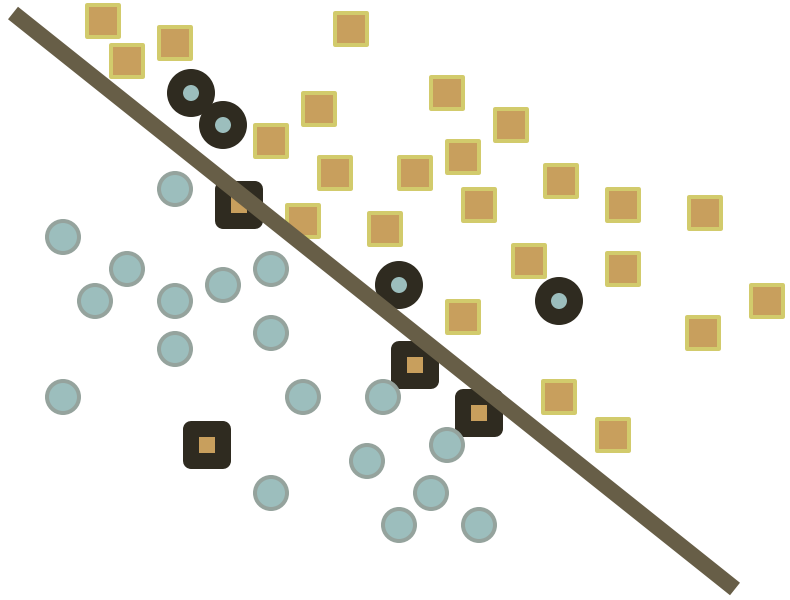
Evaluate Model: Confusion Matrix



isSquare ~ xLocation + yLocation

Evaluate Model : Train Model 1

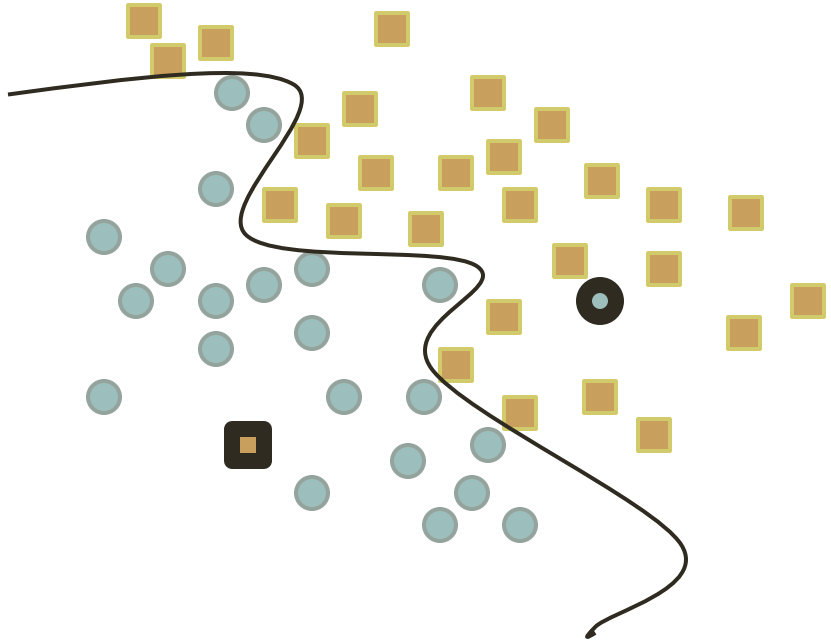
	Sq	Ci
P	36	4
N	4	26



isSquare ~ xLocation + yLocation

Evaluate Model : Train Model 2

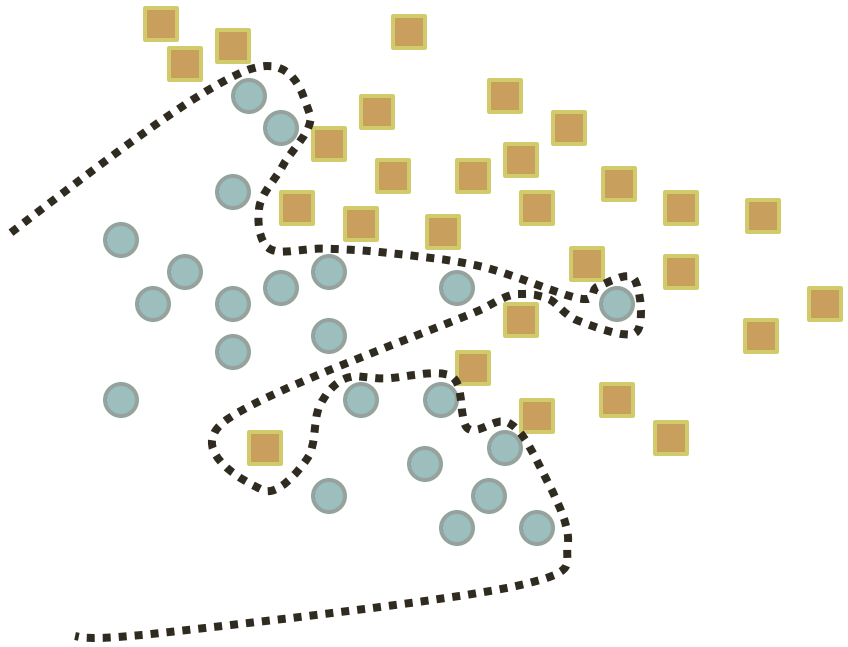
	Sq	Ci
P	39	1
N	1	29



isSquare ~ xLocation + yLocation

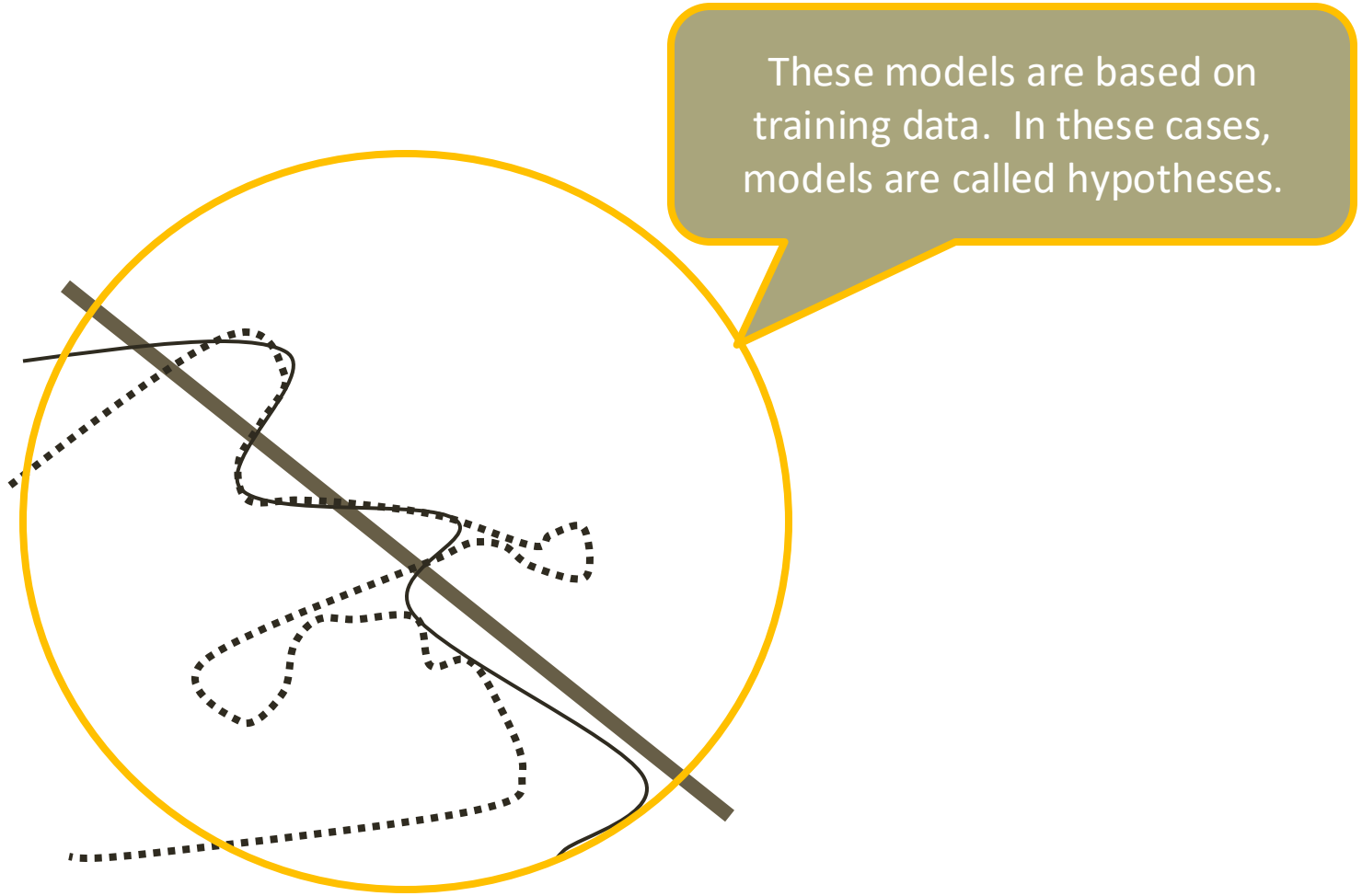
Evaluate Model : Train Model 3

	Sq	Ci
P	40	0
N	0	30



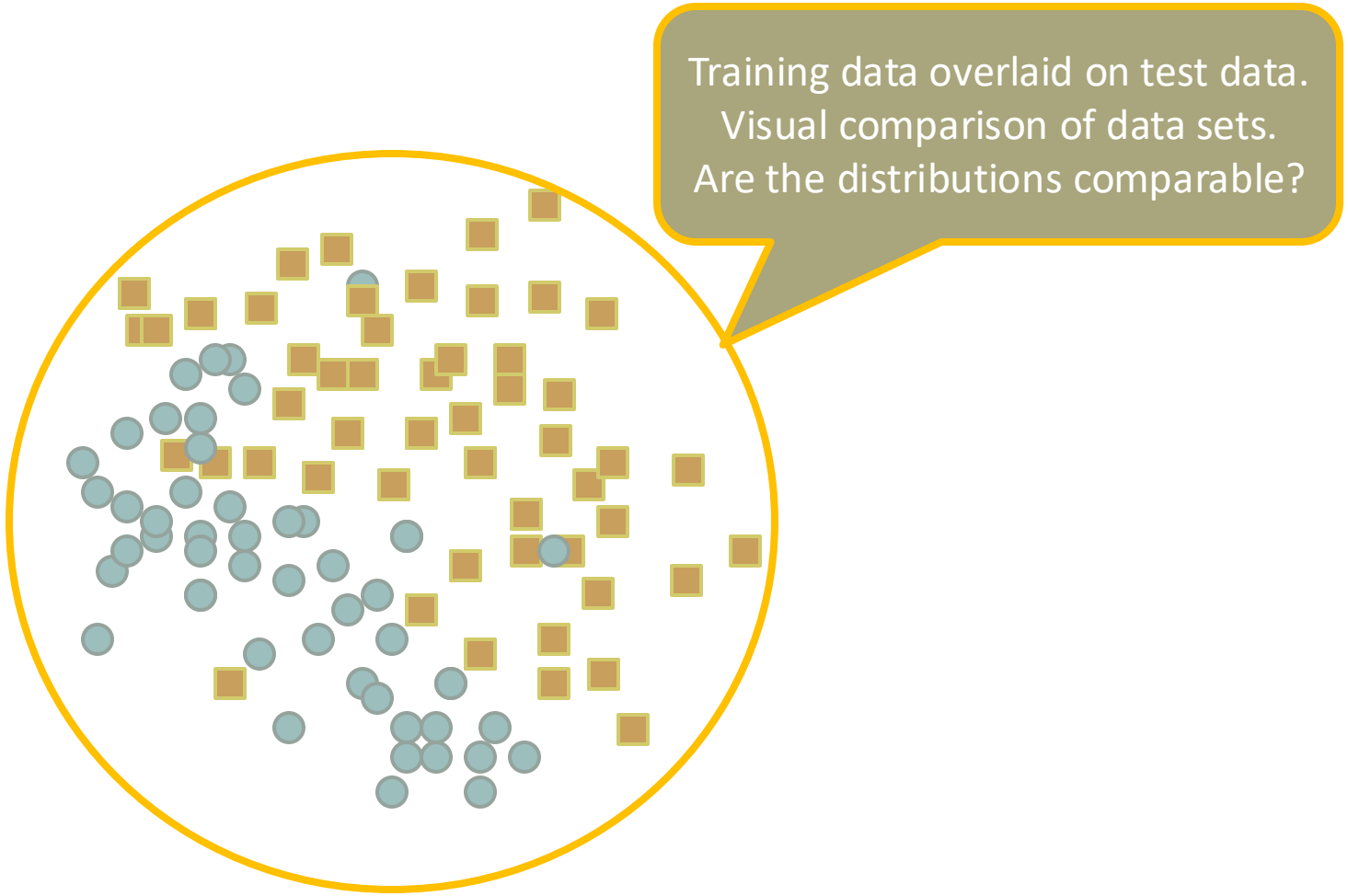
isSquare ~ xLocation + yLocation

Evaluate Model : 3 Models



isSquare ~ xLocation + yLocation

Evaluate Model : All Data

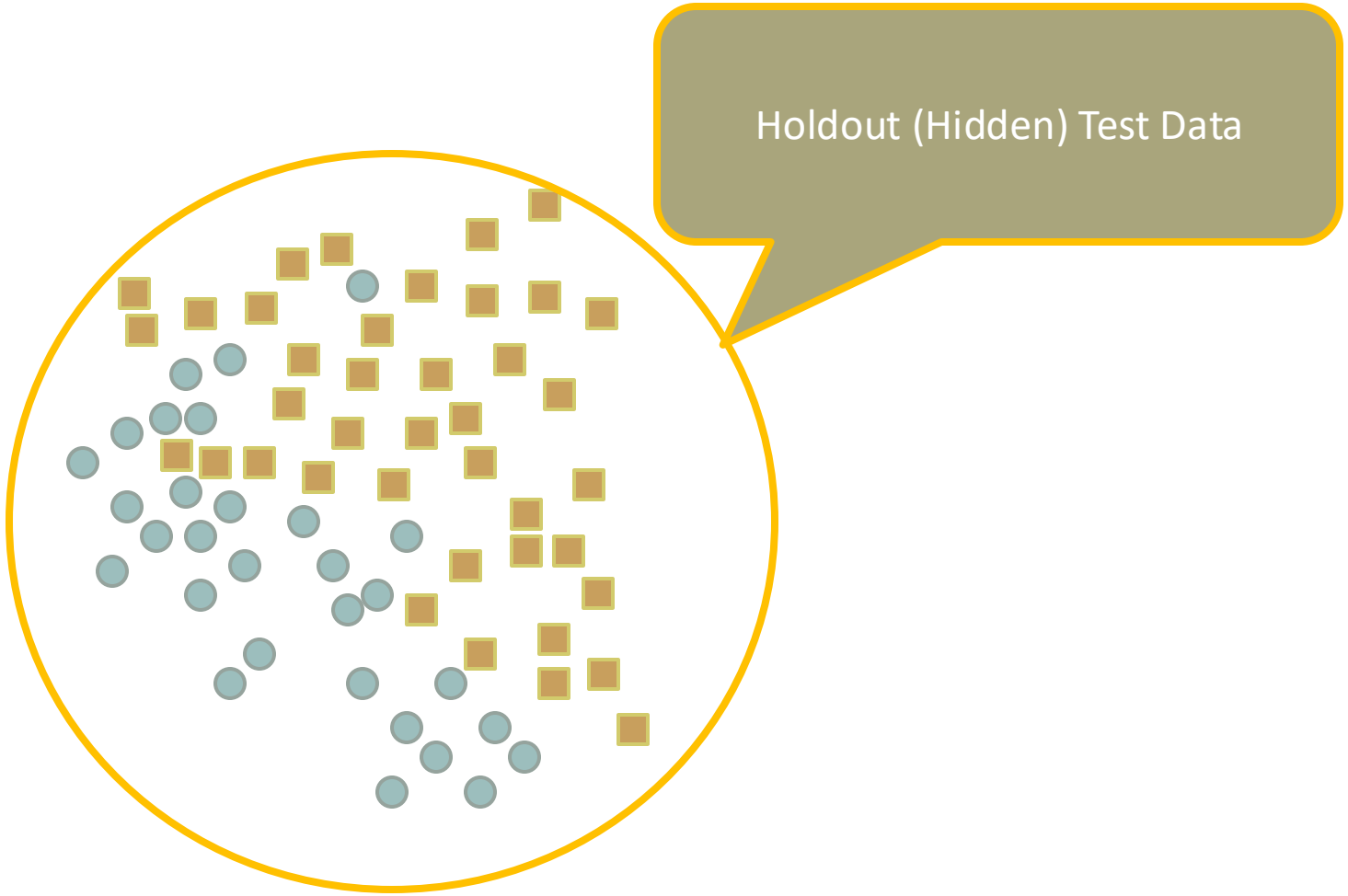


isSquare ~ xLocation + yLocation

Evaluate Model : Training Data

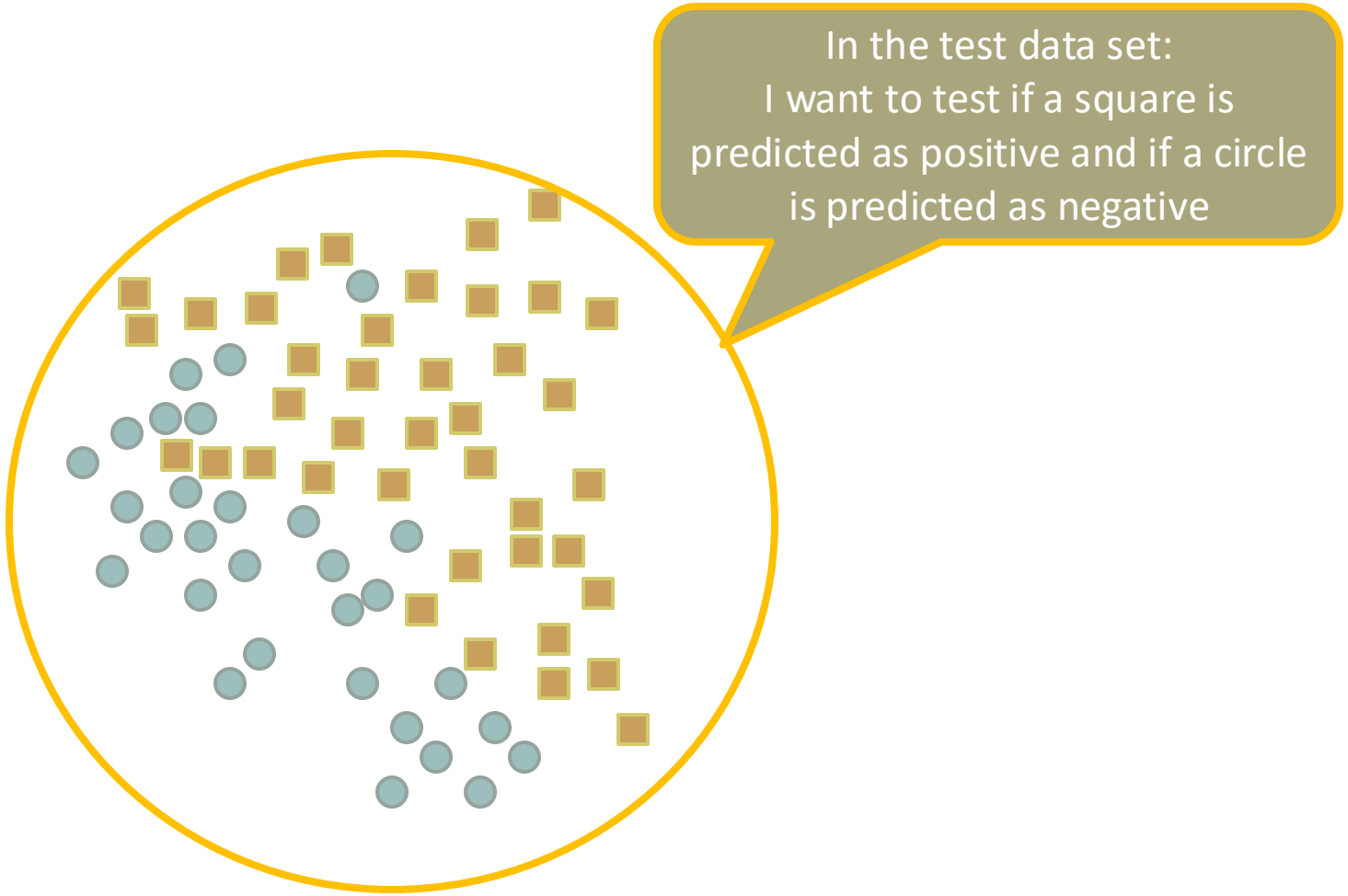
isSquare ~ xLocation + yLocation

Evaluate Model : Test Data



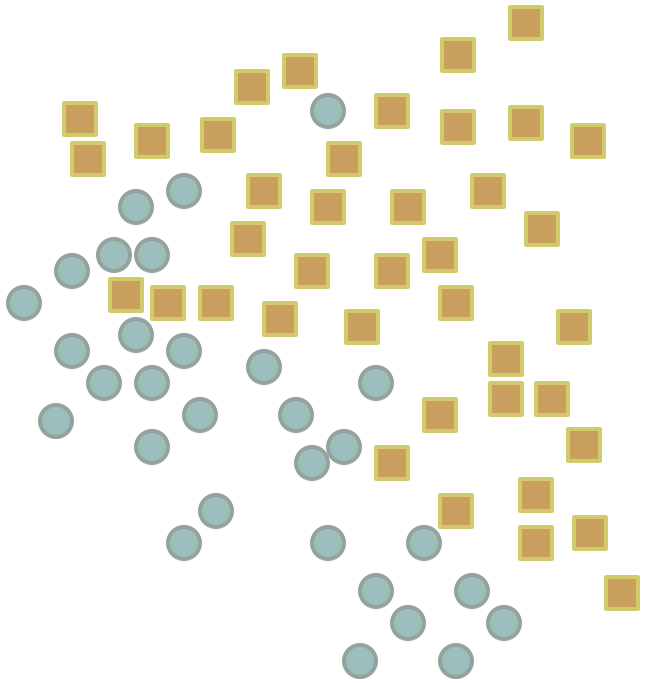
isSquare ~ xLocation + yLocation

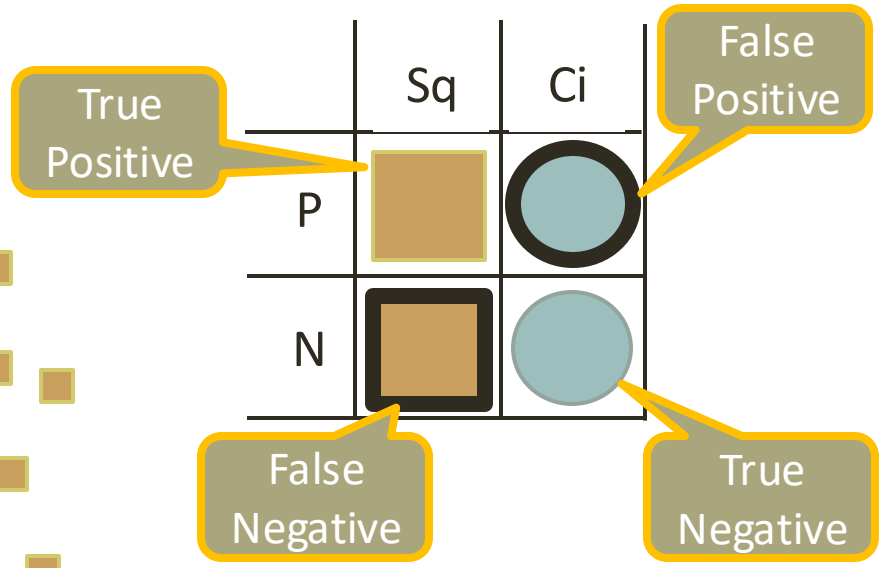
Evaluate Model : Test Data



isSquare ~ xLocation + yLocation

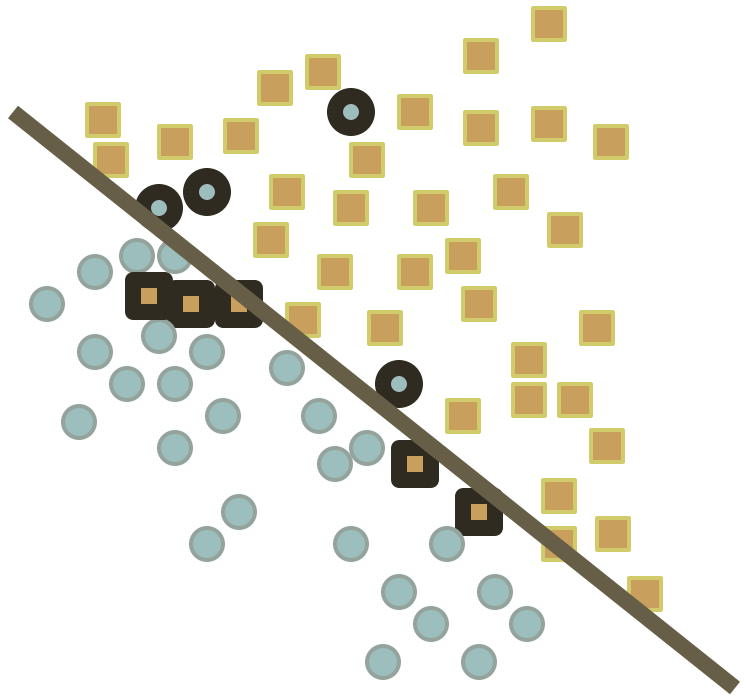
Evaluate Model : Test Data





isSquare ~ xLocation + yLocation

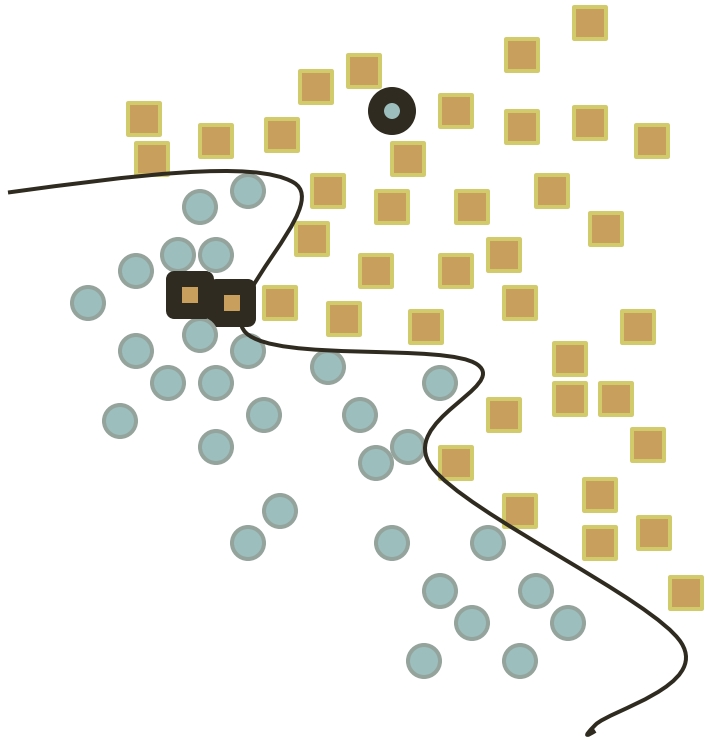
Evaluate Model : Test Model 1



	Sq	Ci
P	35	4
N	5	26

isSquare ~ xLocation + yLocation

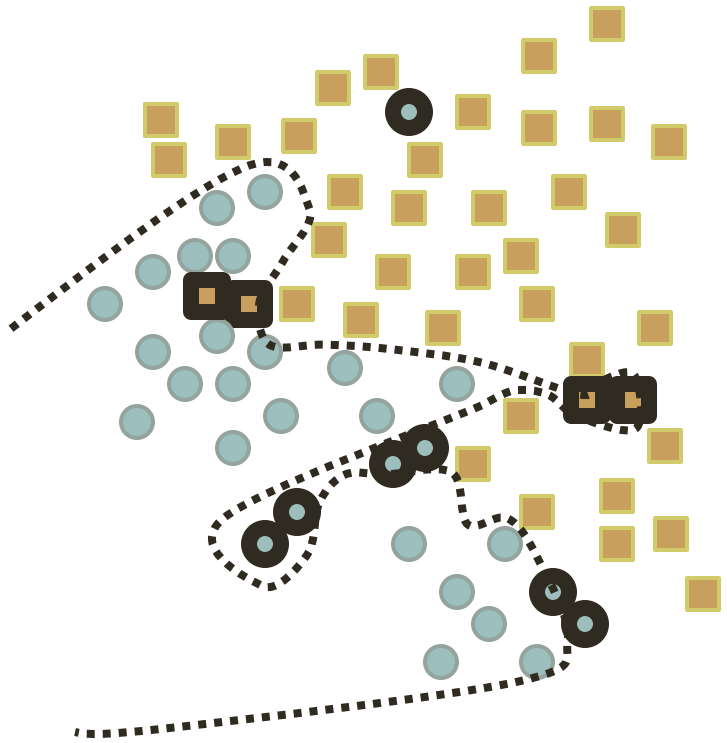
Evaluate Model : Test Model 2



	Sq	Ci
P	38	1
N	2	29

isSquare ~ xLocation + yLocation

Evaluate Model : Test Model 3



	Sq	Ci
P	36	7
N	4	23

isSquare ~ xLocation + yLocation

Over-fitting and Confusion Matrix

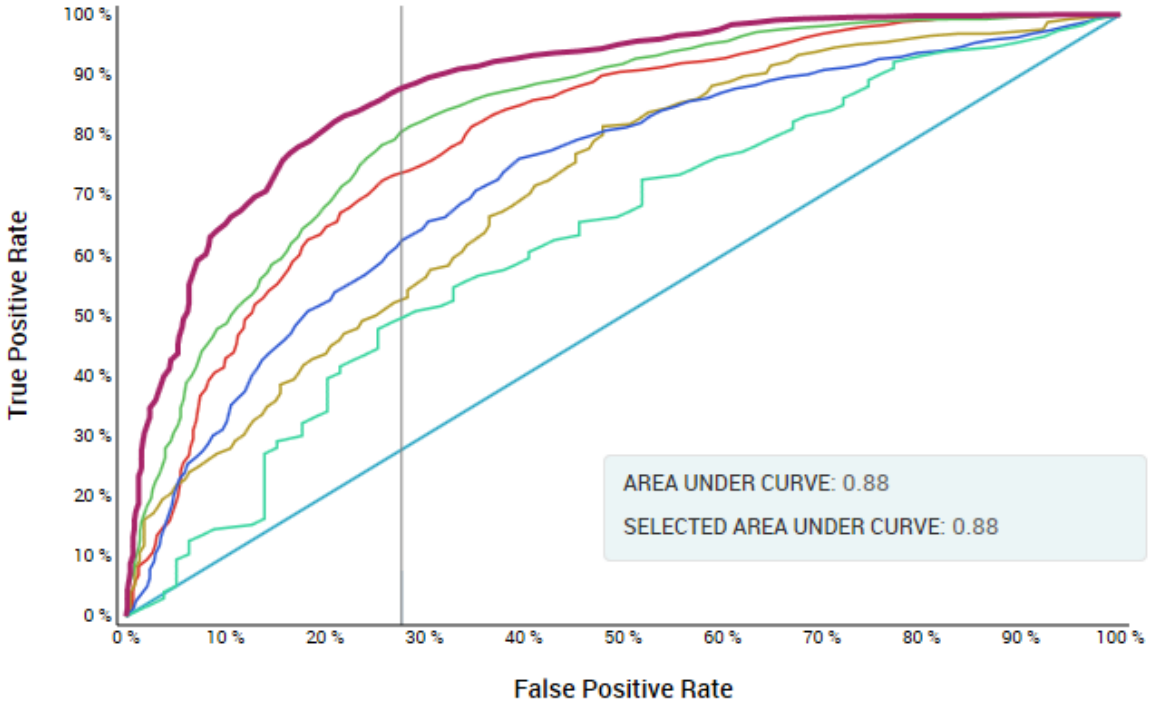
ROC Chart Intro

ROC Chart Intro (1)

- Confusion Matrix and ROC Chart

Generated from test data

Receiver Operating Characteristic Curve



Classification Matrix

Model:C0_R_LogReg

View: Percent | Count

Actual / Predicted	'Does not plan to at...	Not 'Does not plan t...
'Does not plan to at...	1284	207
Not 'Does not plan t...	165	504

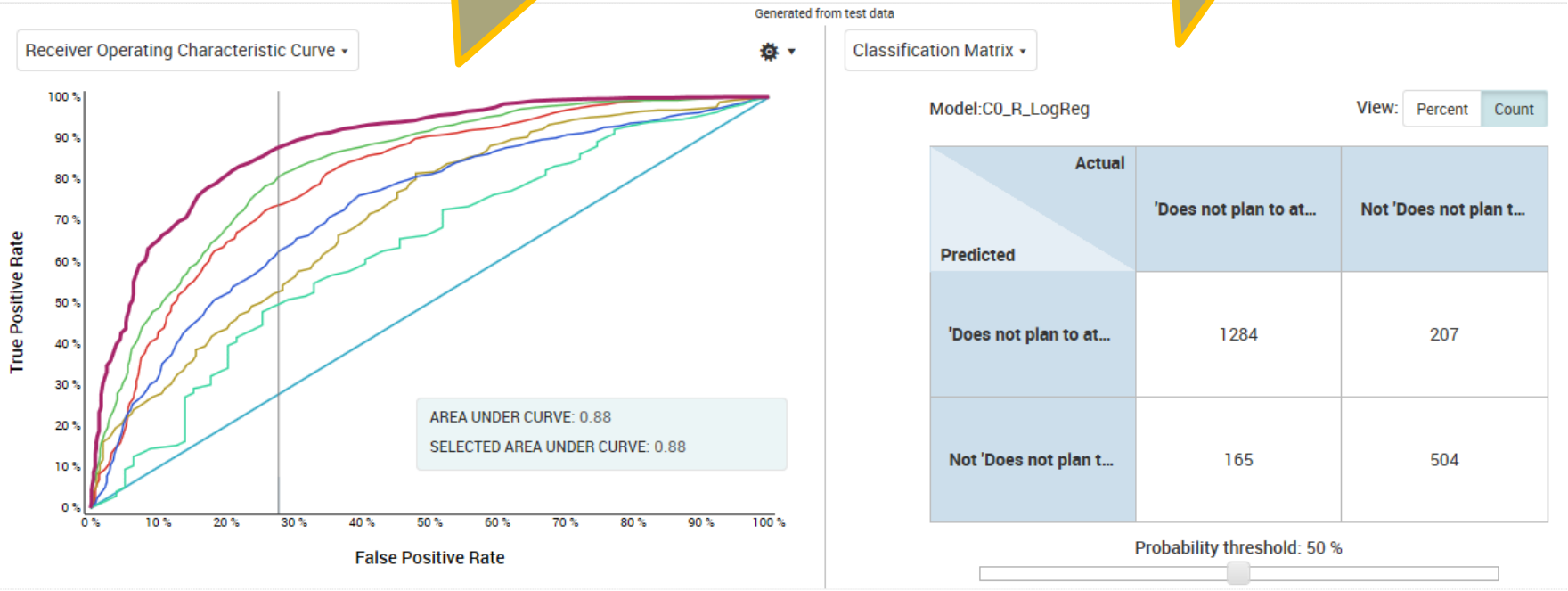
Probability threshold: 50 %

ROC Chart Intro (2)

- Confusion Matrix and ROC Chart

Comparison of 6 ROC curves
Each curve is from a different model

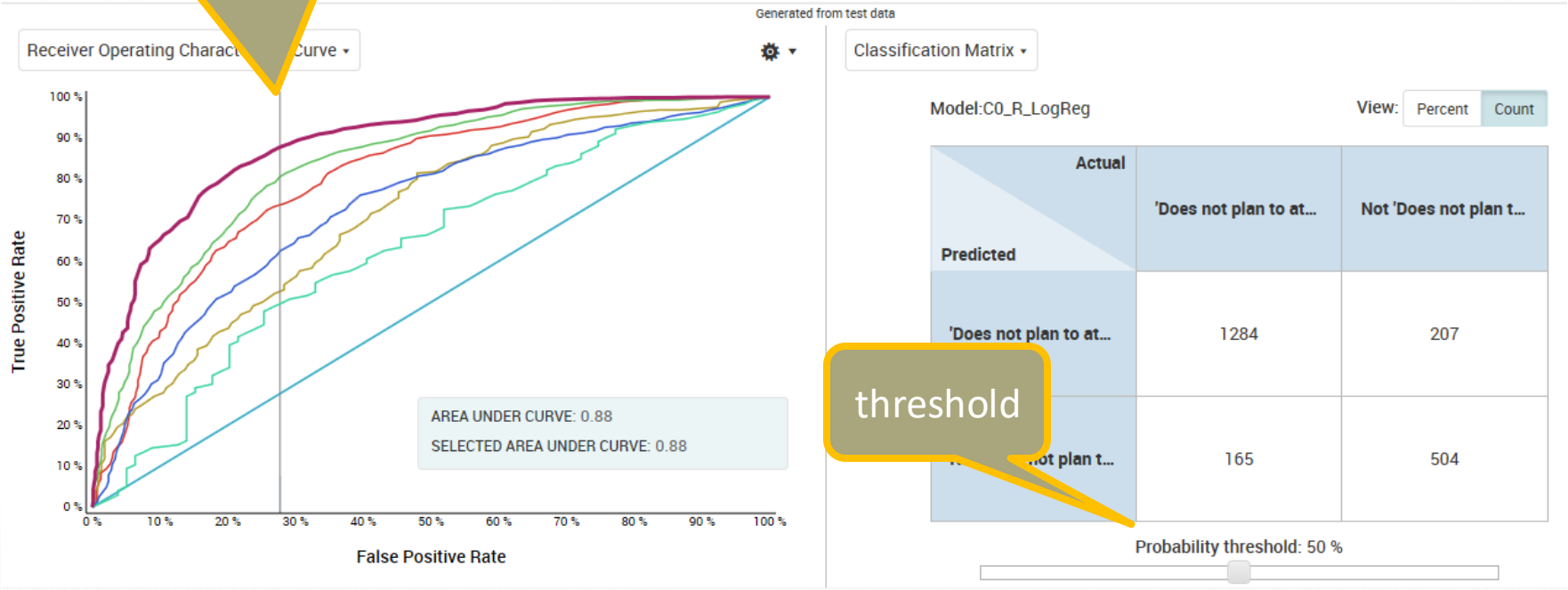
The confusion matrix for one model at one threshold



ROC Chart Intro (3)

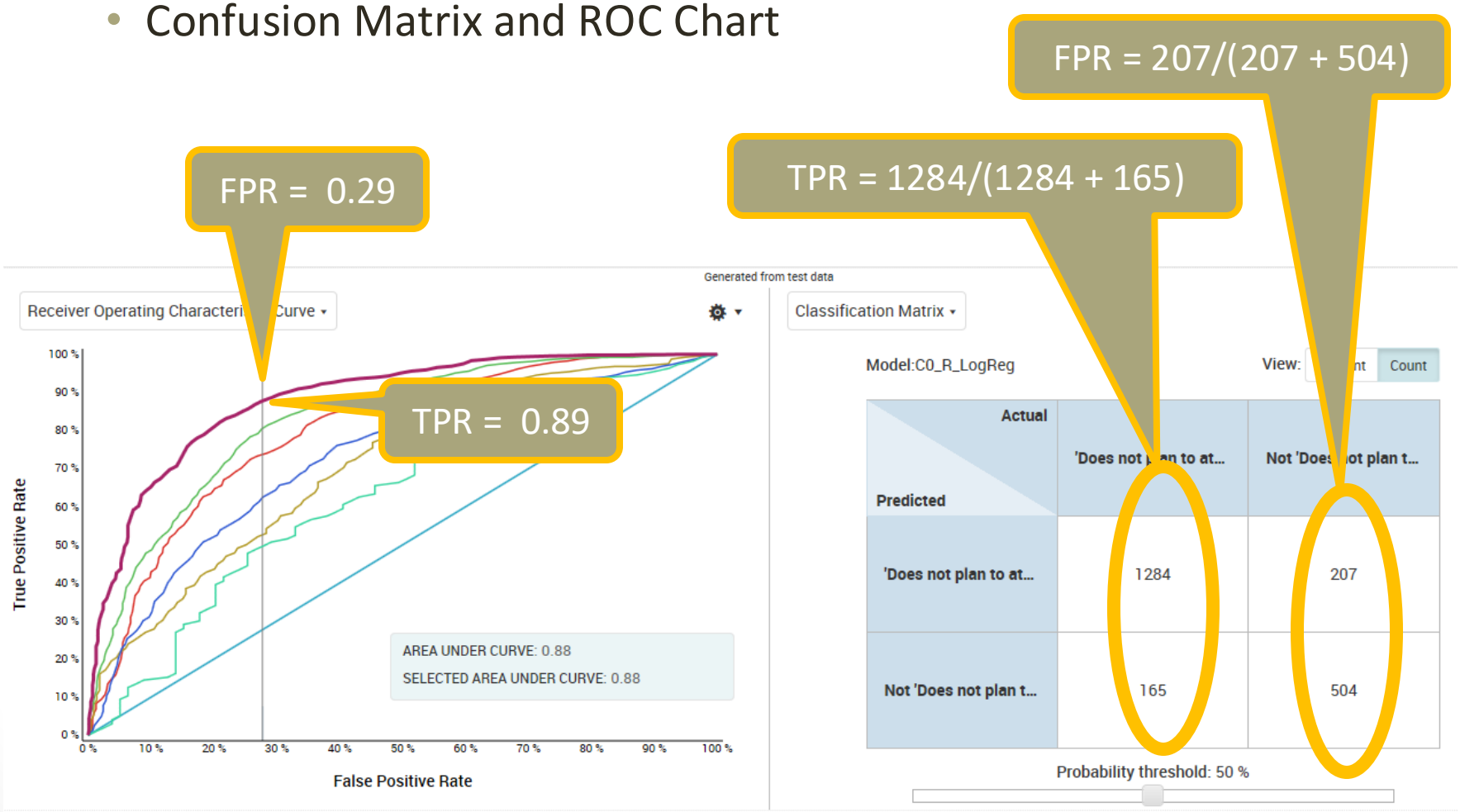
- Confusion Matrix and ROC Chart

This FPR (0.28) corresponds to the threshold (0.5) for the confusion matrix for the best model



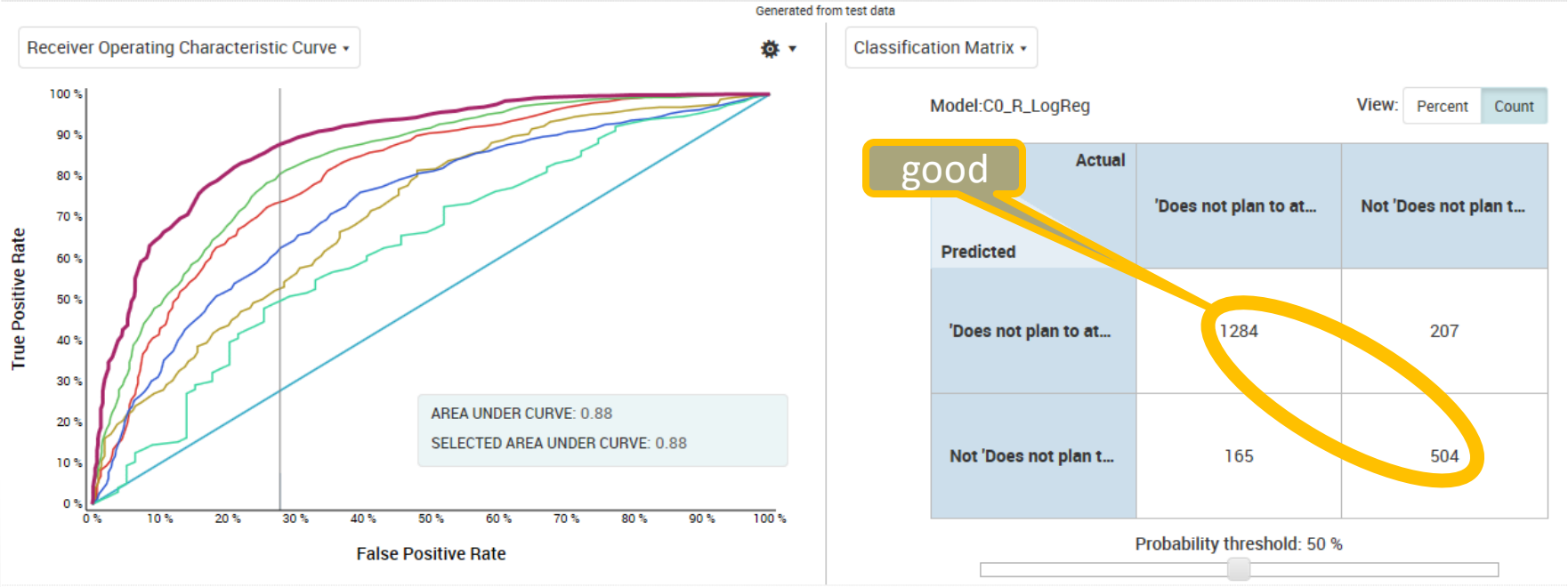
ROC Chart Intro (4)

- Confusion Matrix and ROC Chart



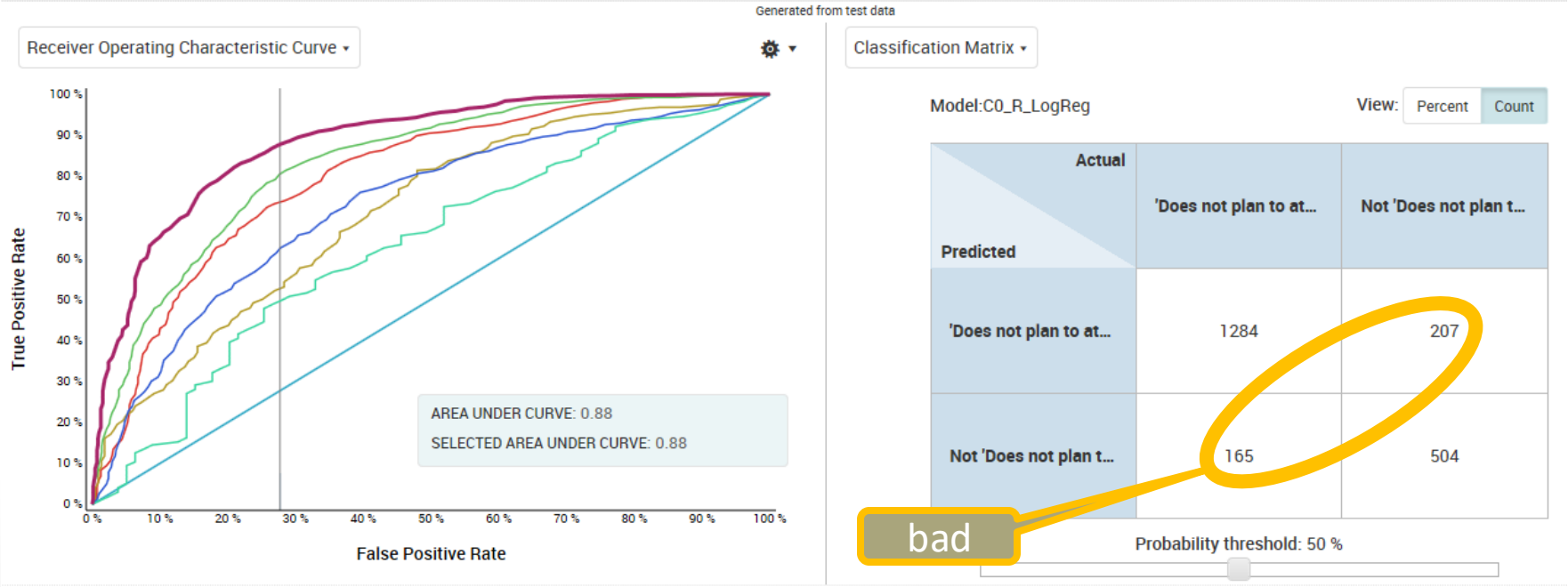
ROC Chart Intro (5)

- Confusion Matrix and ROC Chart



ROC Chart Intro (6)

- Confusion Matrix and ROC Chart

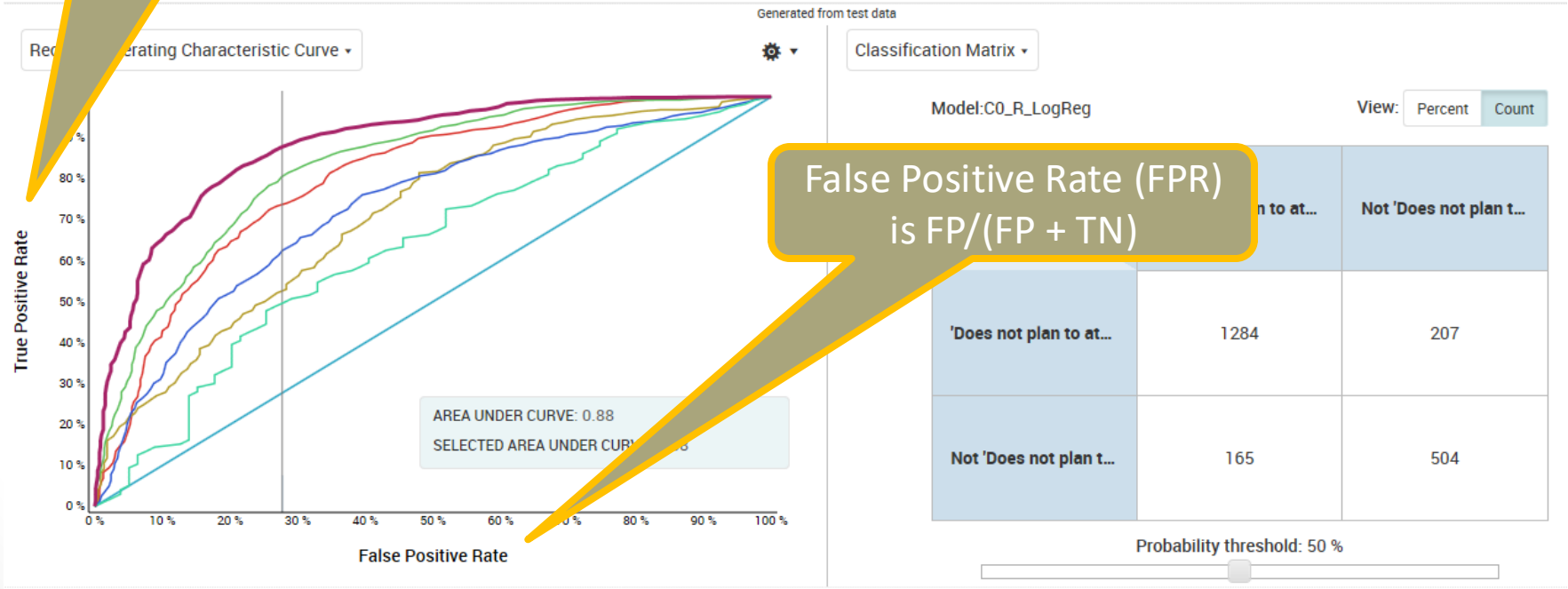


ROC Chart Intro (7)

- Confusion Matrix and ROC Chart

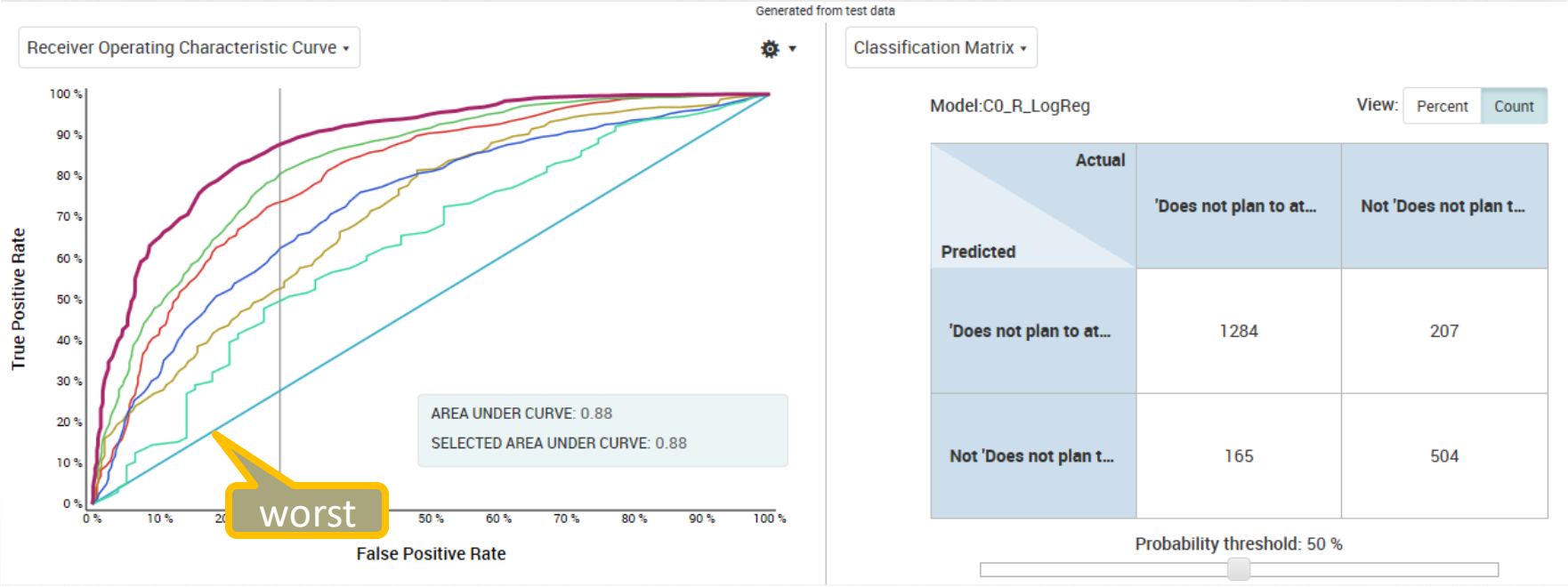
True Positive Rate (TPR) is TP/(TP + FN)

False Positive Rate (FPR) is FP/(FP + TN)



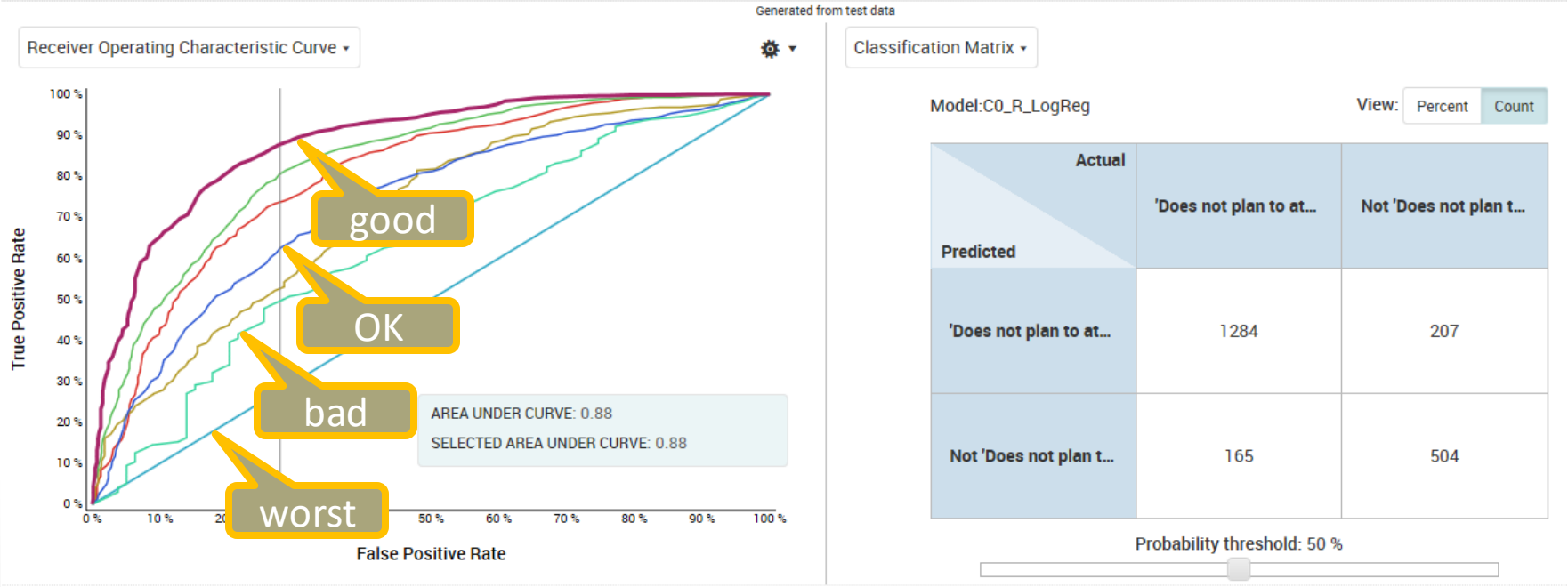
ROC Chart Intro (8)

- Confusion Matrix and ROC Chart



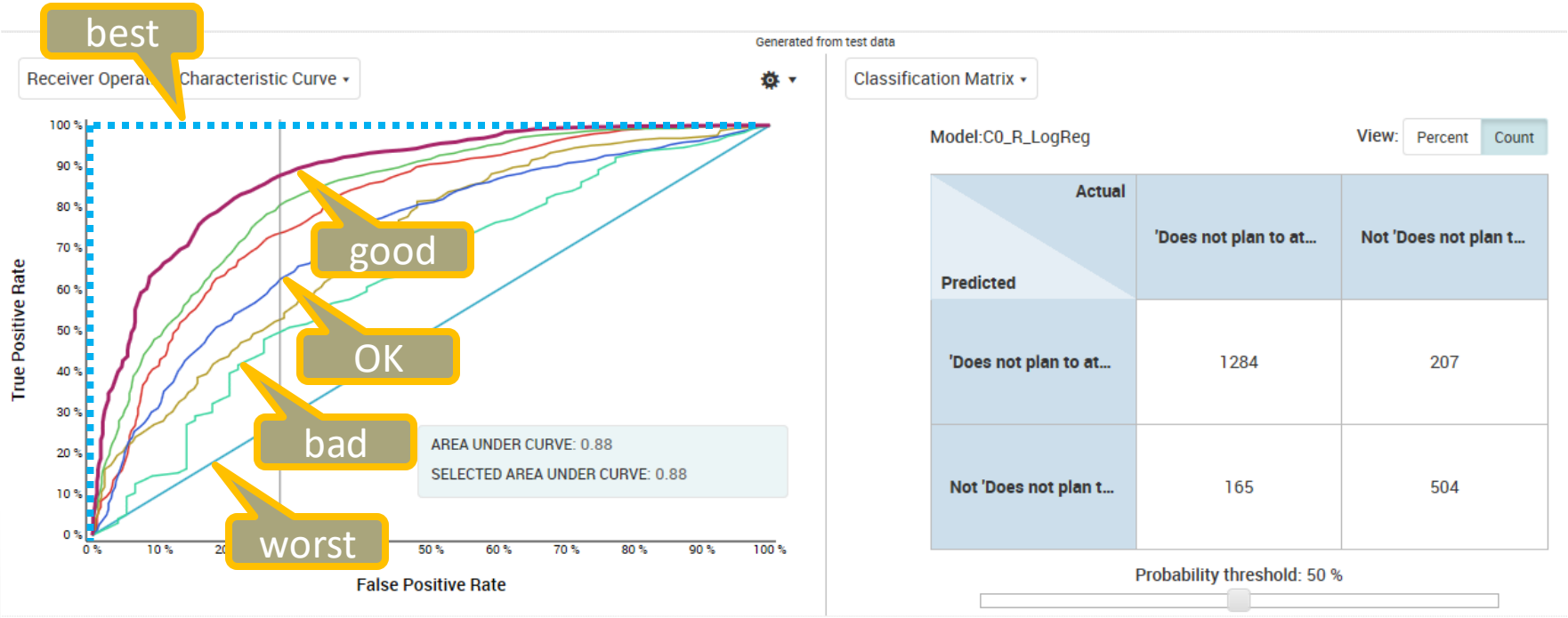
ROC Chart Intro (9)

- Confusion Matrix and ROC Chart



ROC Chart Intro (10)

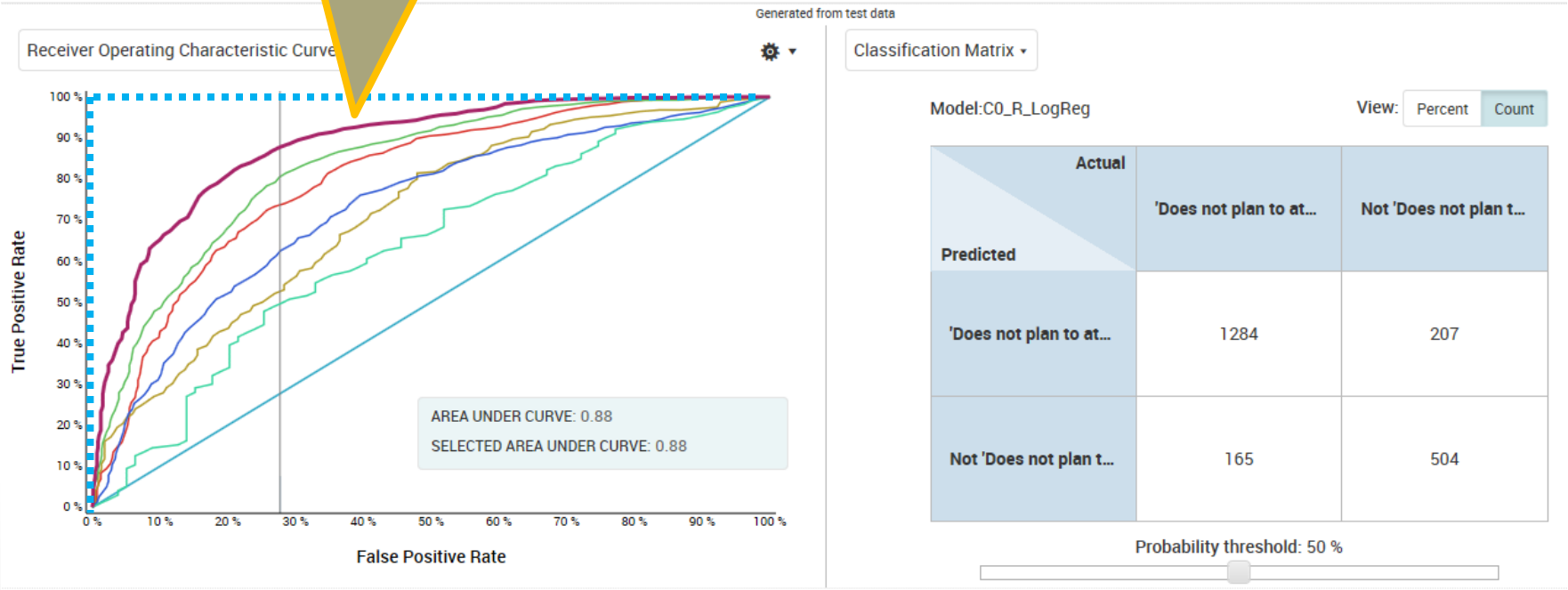
- Confusion Matrix and ROC Chart



ROC Chart Intro (11)

- Confusion Matrix and ROC Chart

ROC charts are non-decreasing functions



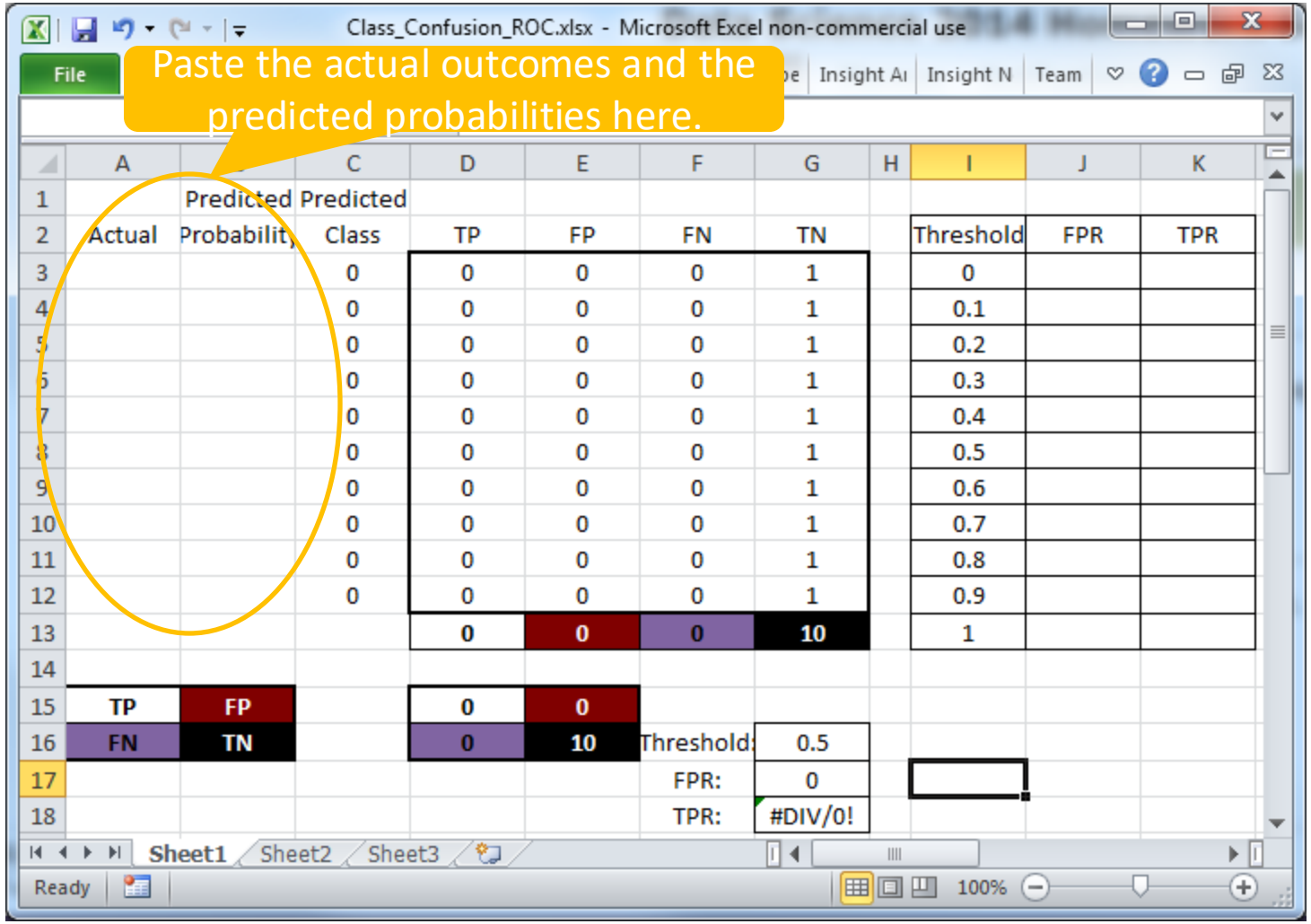
ROC Chart Intro

How to make an ROC

How to make an ROC (0)

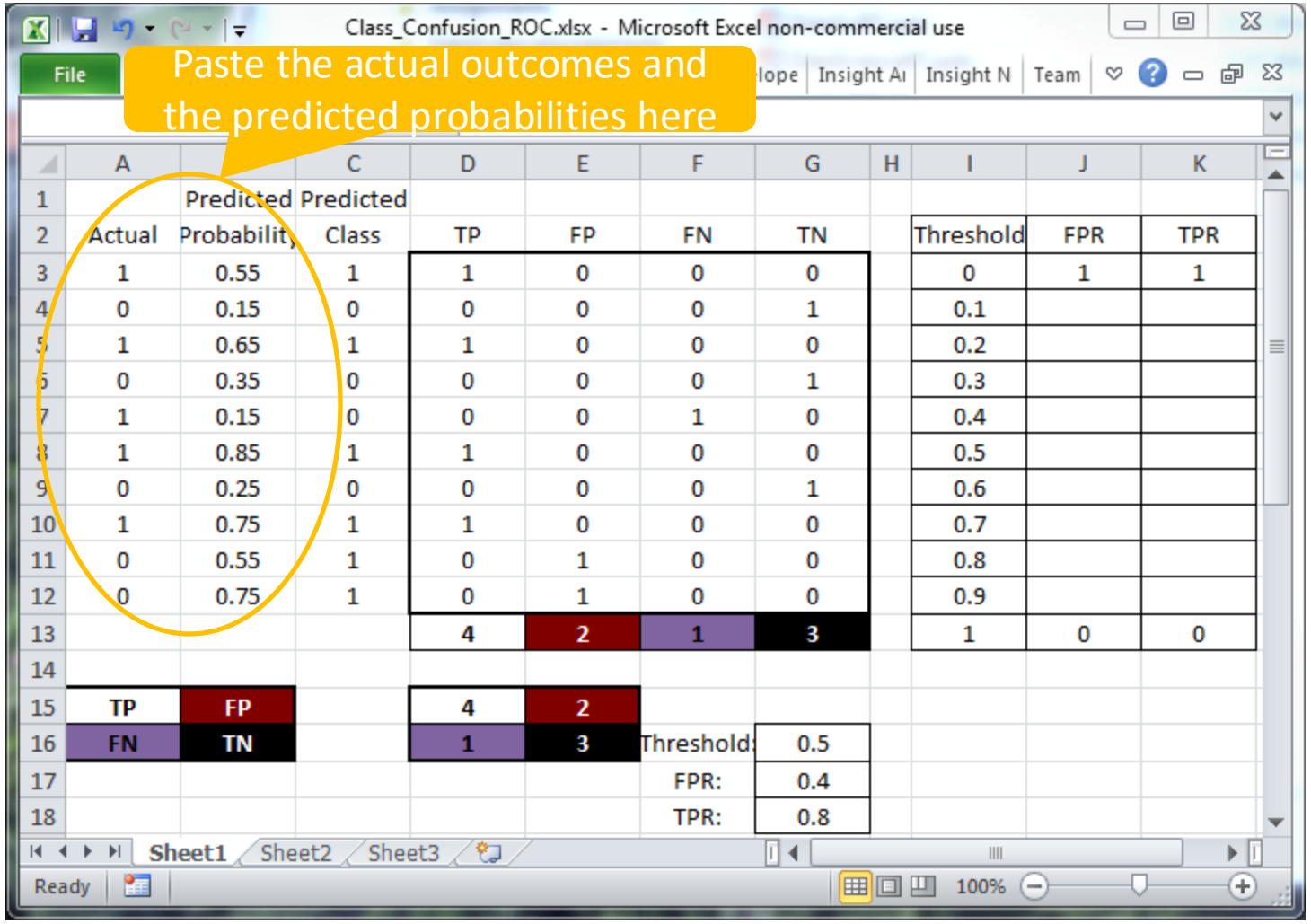
- From Probabilities to ROC:
- Probabilities -> Threshold -> Predictions -> Confusion Matrix -> ROC
- Get Excel workbook: HowToMakeAnROC.xls
- Note that at the bottom of the worksheet are the actual outcomes and the predicted probabilities.

How to make an ROC (1)



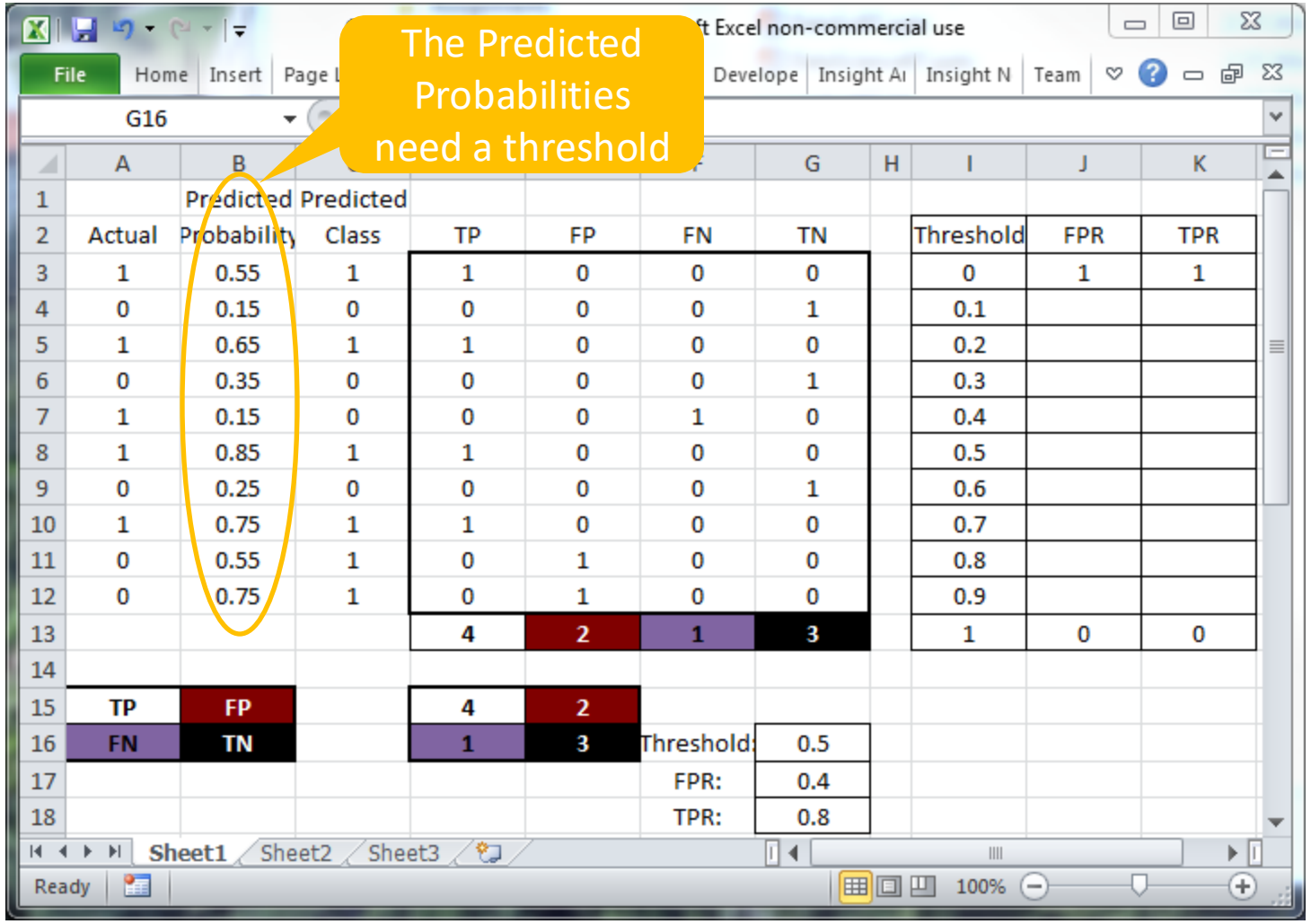
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3			0	0	0	0	1		0		
4			0	0	0	0	1		0.1		
5			0	0	0	0	1		0.2		
6			0	0	0	0	1		0.3		
7			0	0	0	0	1		0.4		
8			0	0	0	0	1		0.5		
9			0	0	0	0	1		0.6		
10			0	0	0	0	1		0.7		
11			0	0	0	0	1		0.8		
12			0	0	0	0	1		0.9		
13				0	0	0	10		1		
14											
15	TP	FP		0	0						
16	FN	TN		0	10	Threshold:	0.5				
17						FPR:	0				
18						TPR:	#DIV/0!				

How to make an ROC (2)



	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

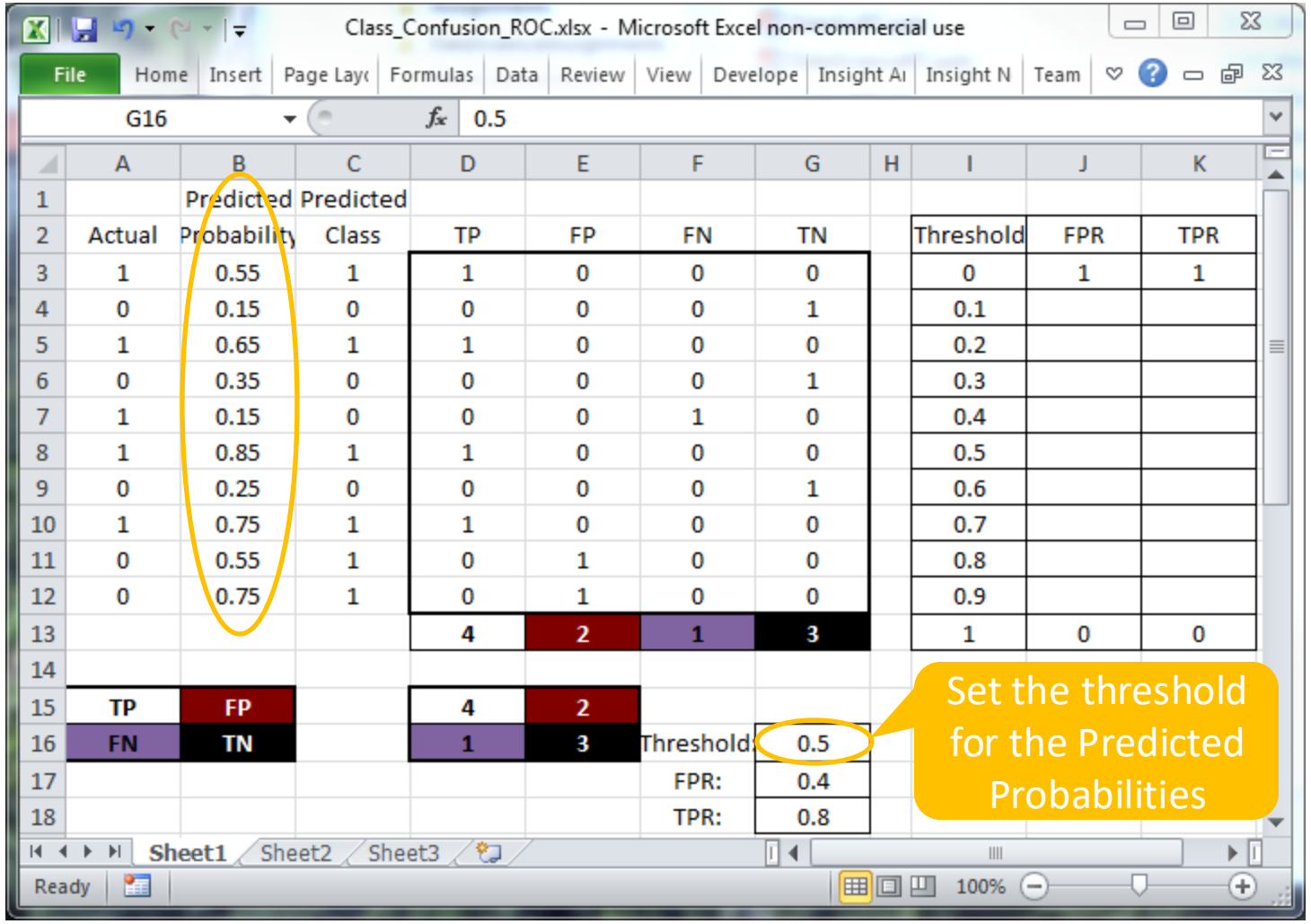
How to make an ROC (3)



	A	B	C	G	H	I	J	K
1		Predicted	Predicted					
2	Actual	Probability	Class	TP	FP	FN	TN	
3	1	0.55	1	1	0	0	0	
4	0	0.15	0	0	0	0	1	
5	1	0.65	1	1	0	0	0	
6	0	0.35	0	0	0	0	1	
7	1	0.15	0	0	0	1	0	
8	1	0.85	1	1	0	0	0	
9	0	0.25	0	0	0	0	1	
10	1	0.75	1	1	0	0	0	
11	0	0.55	1	0	1	0	0	
12	0	0.75	1	0	1	0	0	
13				4	2	1	3	
14								
15	TP	FP		4	2			
16	FN	TN		1	3	Threshold:	0.5	
17						FPR:	0.4	
18						TPR:	0.8	

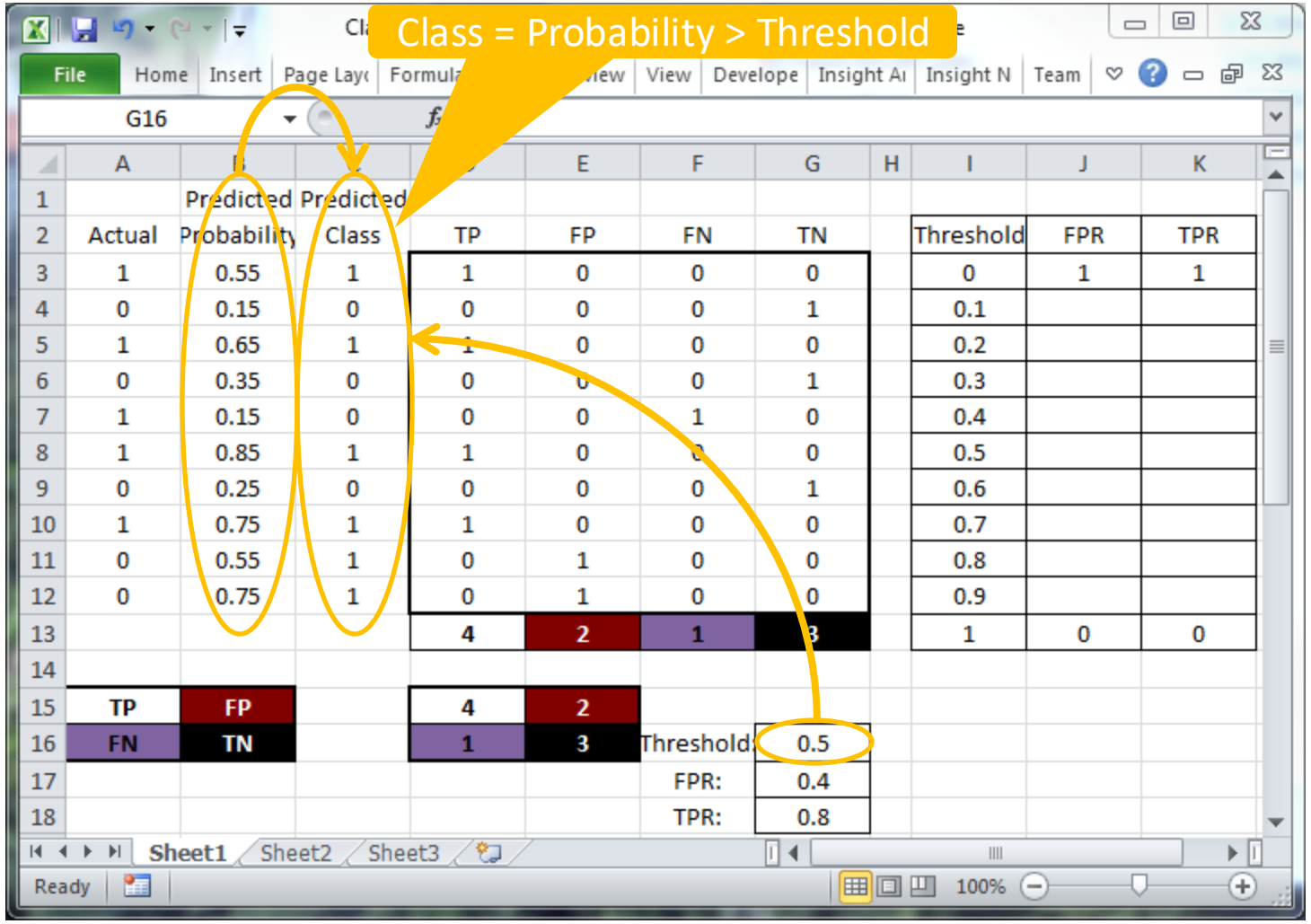
Threshold	FPR	TPR
0	1	1
0.1		
0.2		
0.3		
0.4		
0.5		
0.6		
0.7		
0.8		
0.9		
1	0	0

How to make an ROC (4)



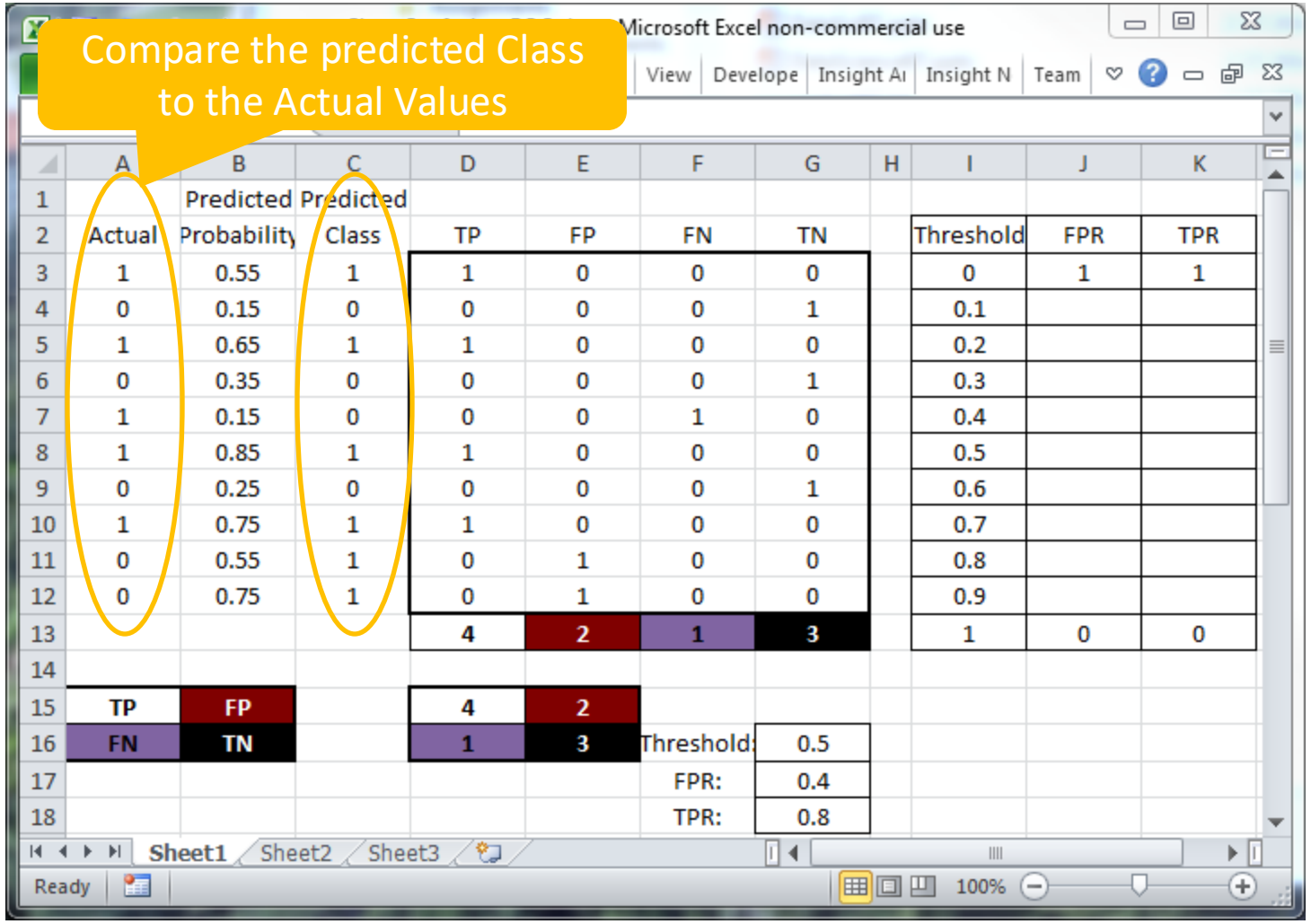
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

How to make an ROC (5)



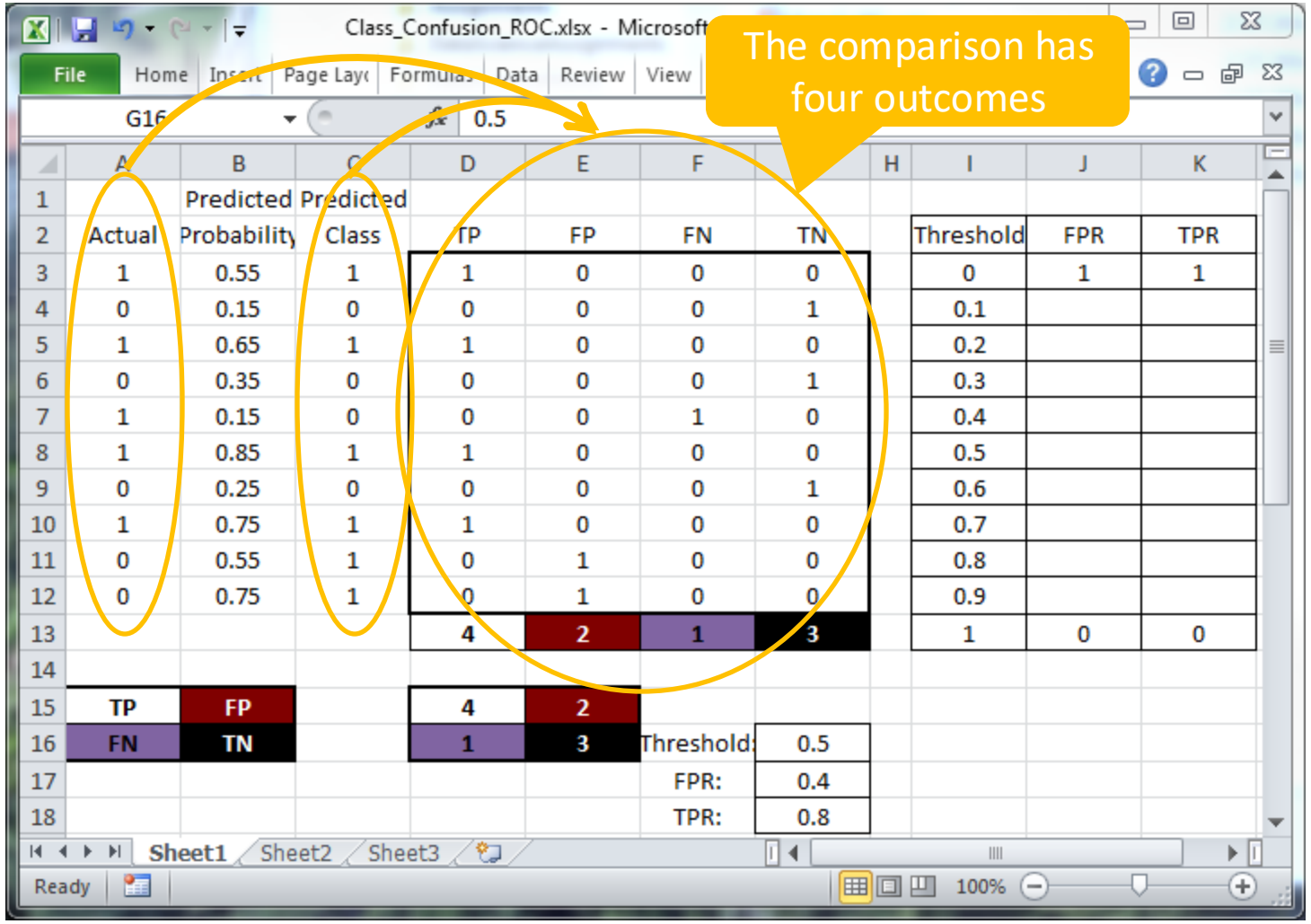
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

How to make an ROC (6)



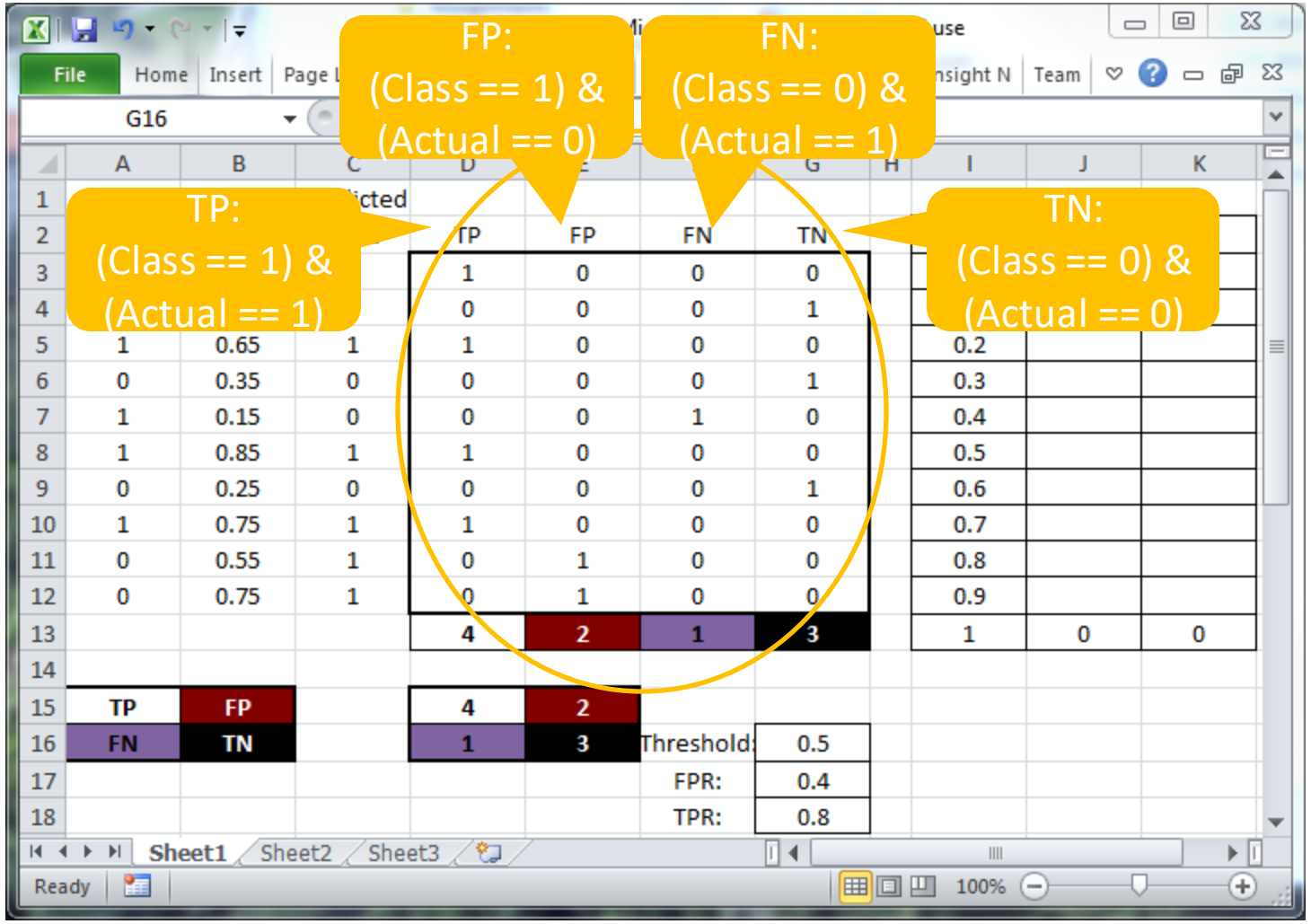
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

How to make an ROC (7)

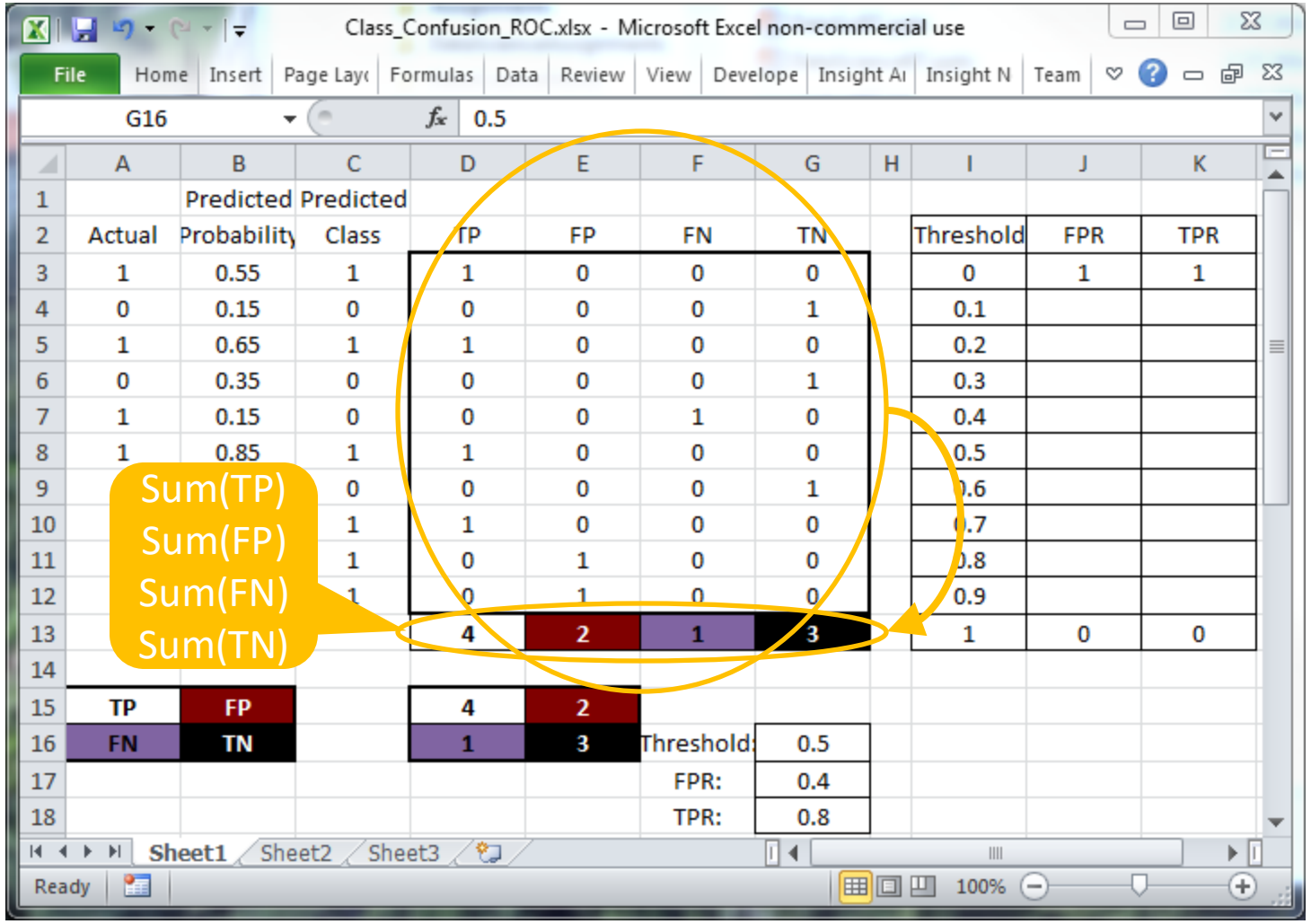


	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

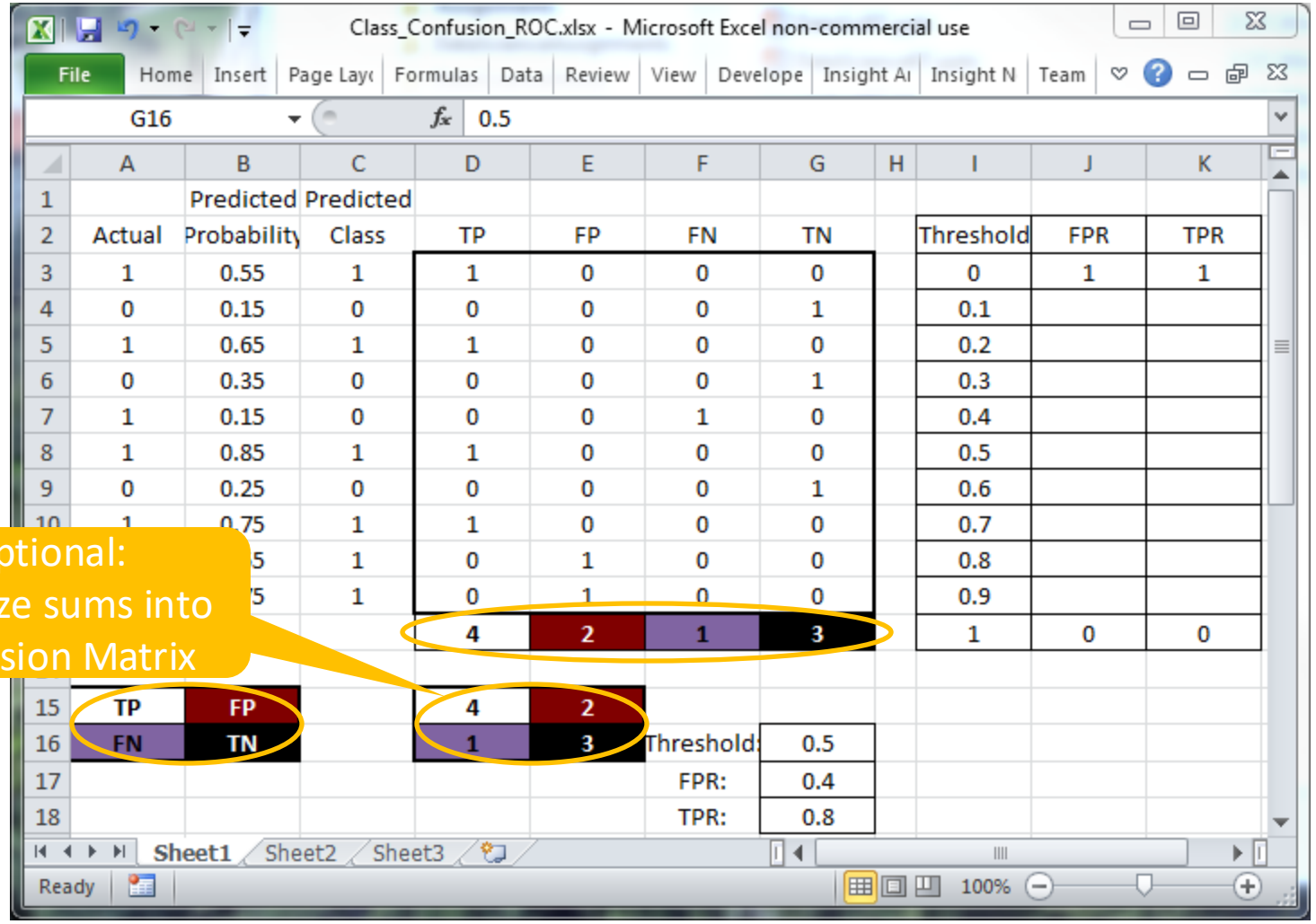
How to make an ROC (8)



How to make an ROC (9)



How to make an ROC (10)



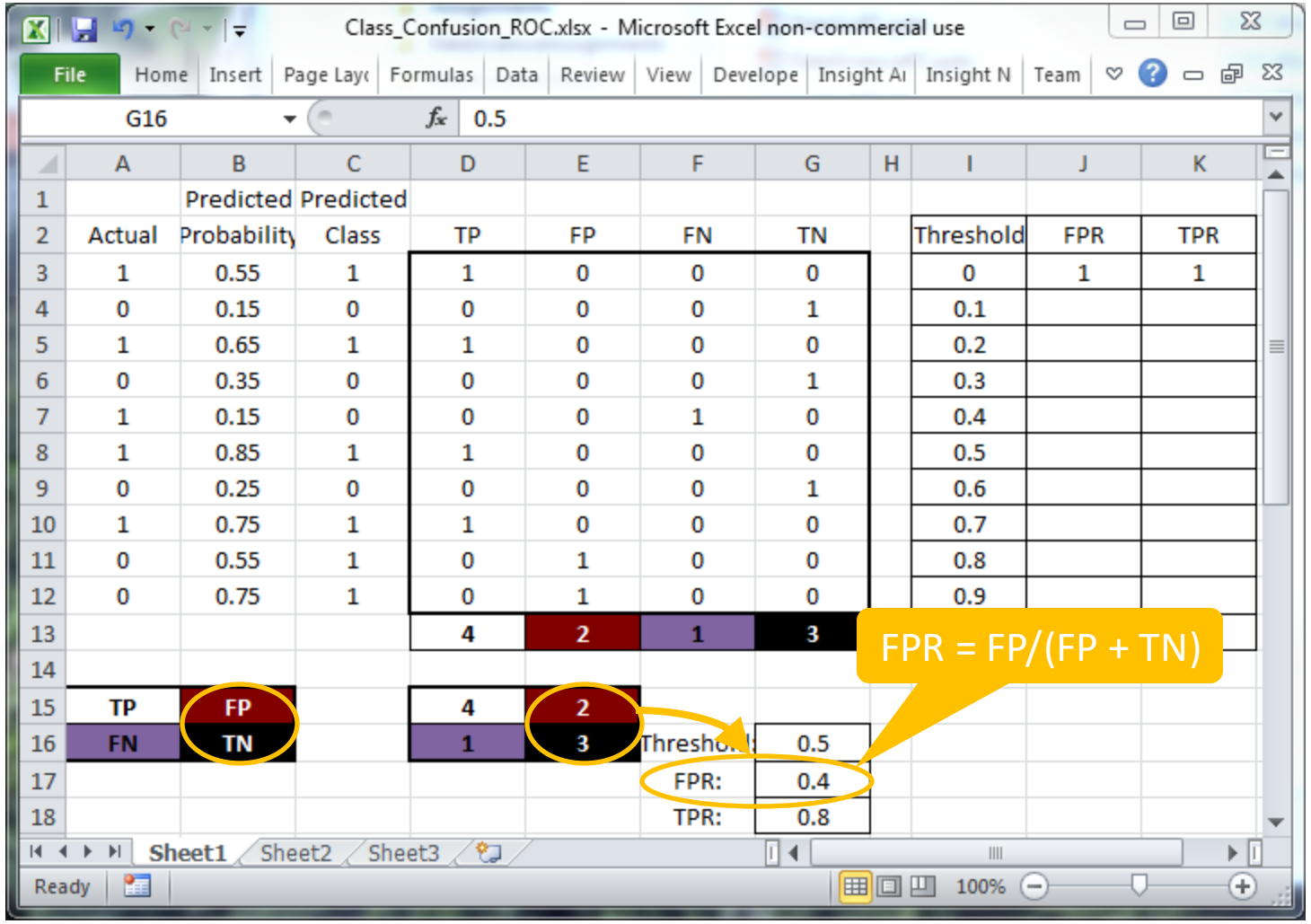
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
		5	1	0	1	0	0		0.8		
		5	1	0	1	0	0		0.9		
				4	2	1	3		1	0	0

	TP	FP
	FN	TN

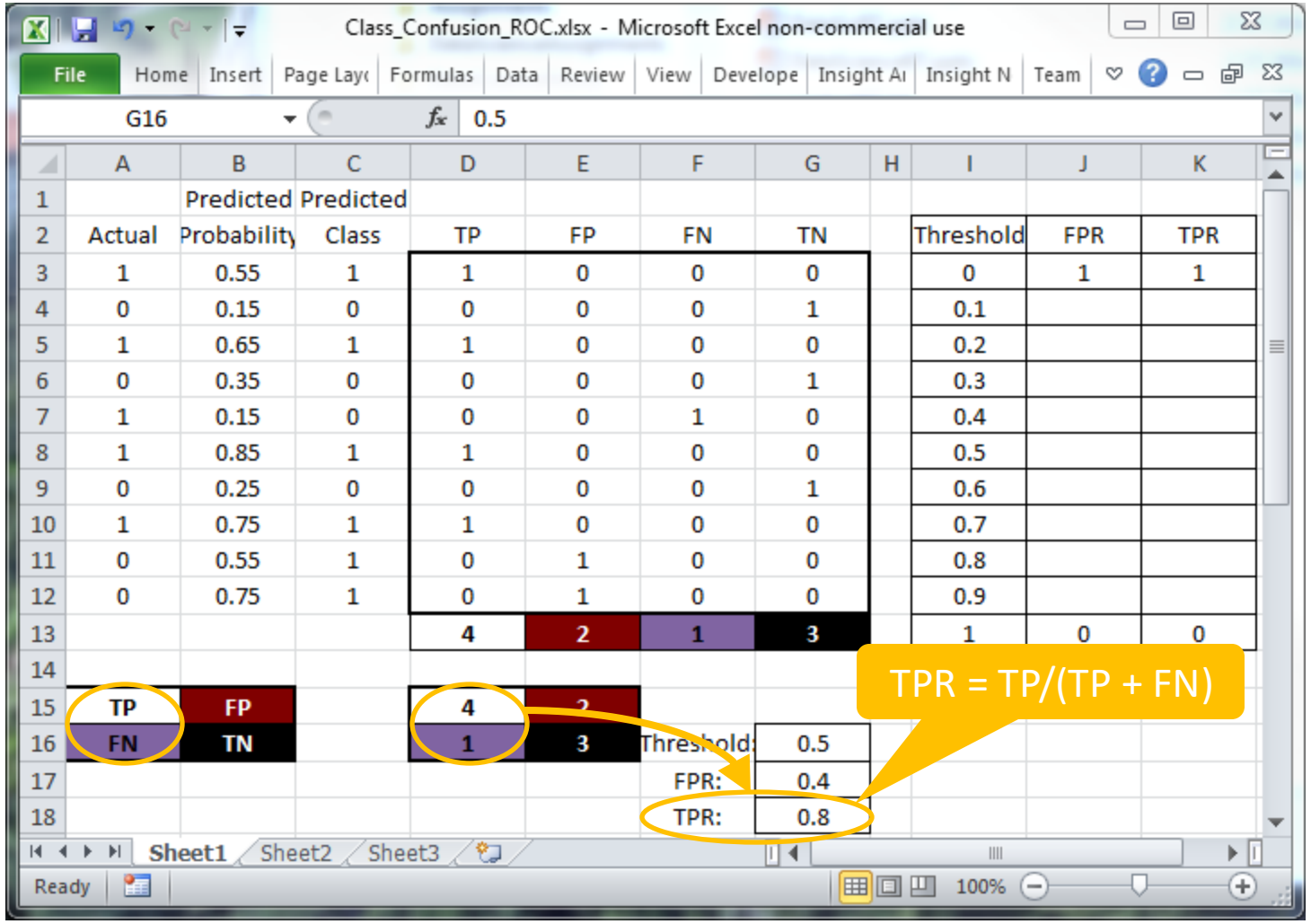
4	2
1	3

Threshold:	0.5
FPR:	0.4
TPR:	0.8

How to make an ROC (11)

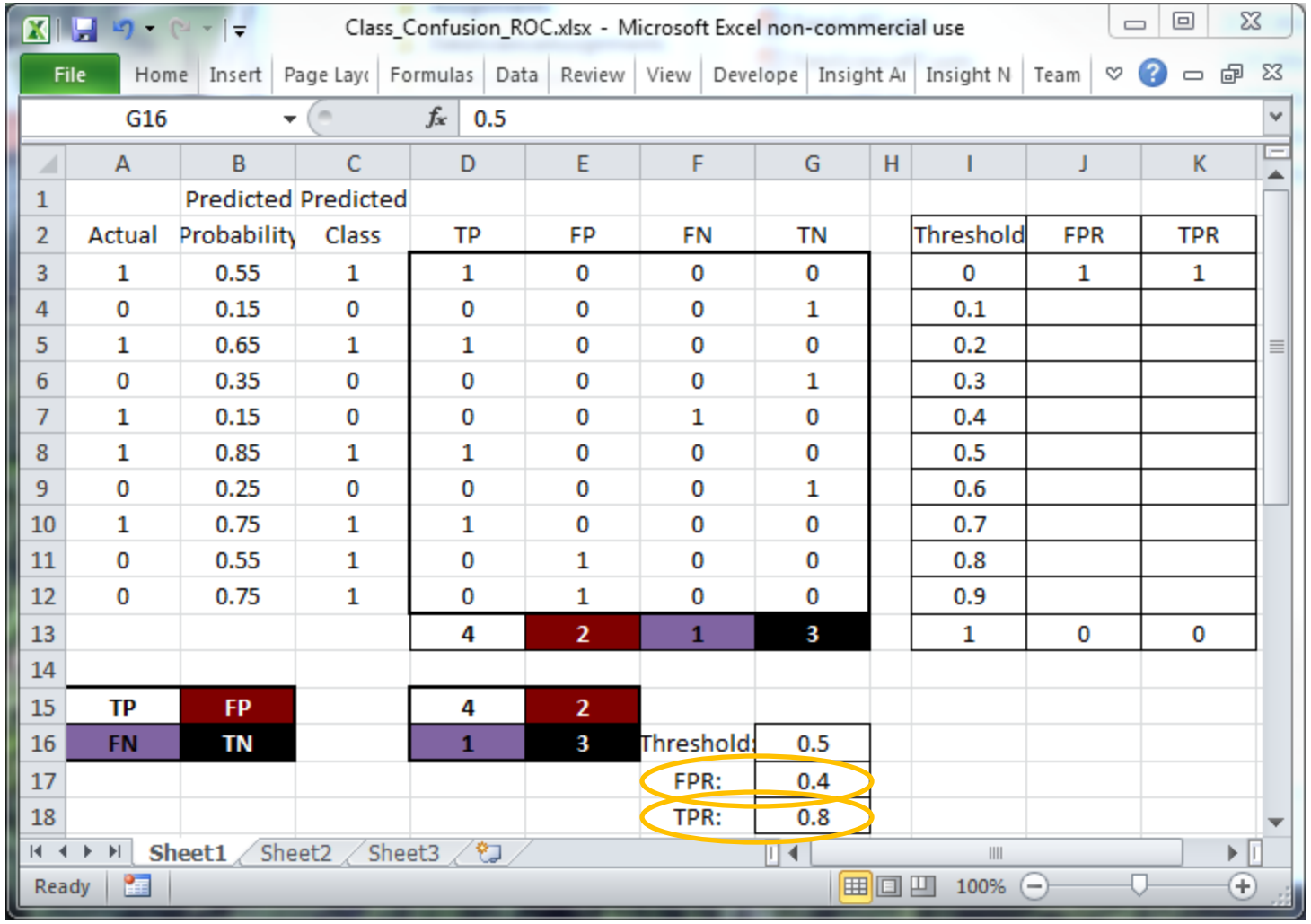


How to make an ROC (12)



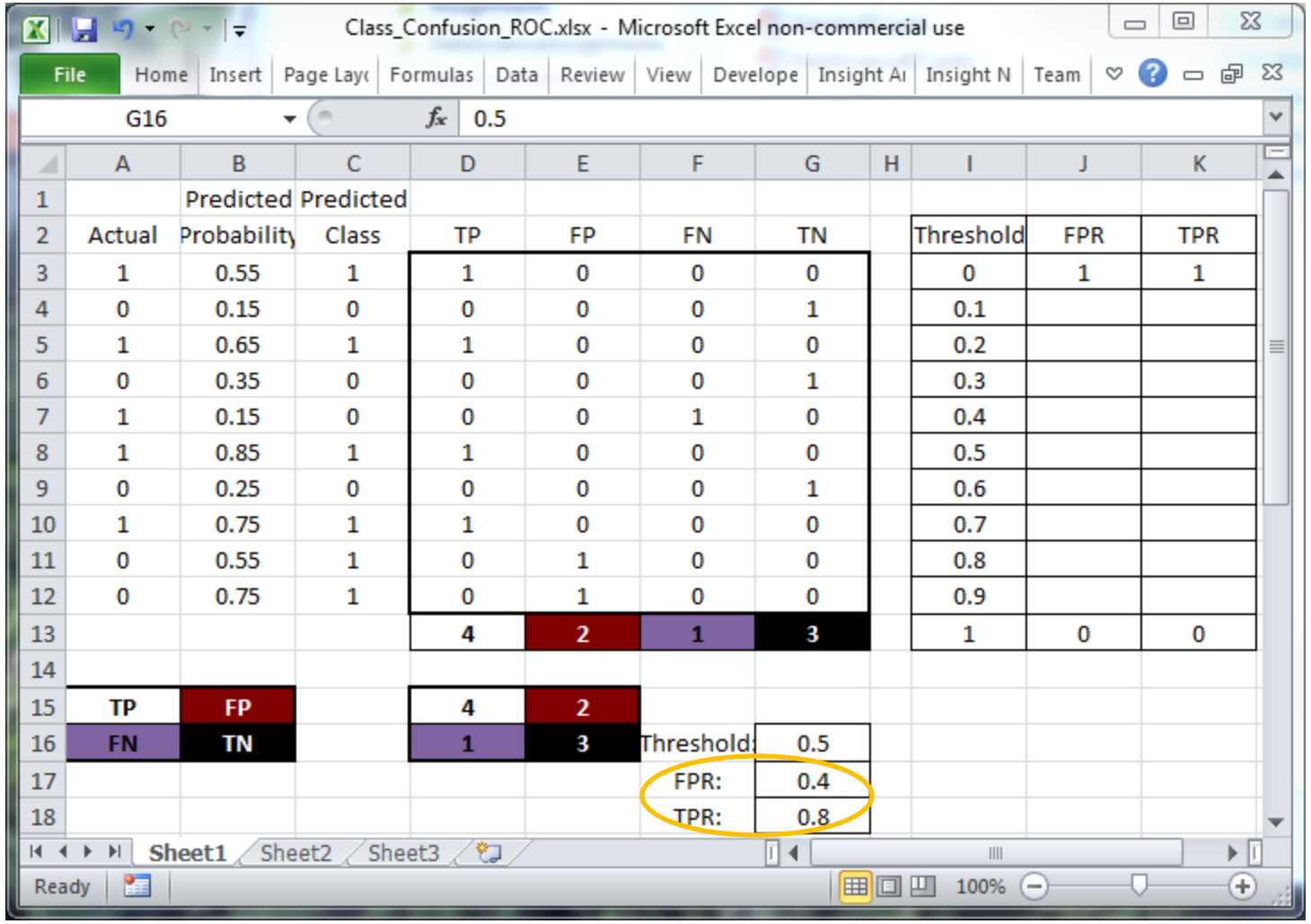
	A	B	C	D	E	F	G	I	J	K
1		Predicted	Predicted							
2	Actual	Probability	Class	TP	FP	FN	TN	Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0	0	1	1
4	0	0.15	0	0	0	0	1	0.1		
5	1	0.65	1	1	0	0	0	0.2		
6	0	0.35	0	0	0	0	1	0.3		
7	1	0.15	0	0	0	1	0	0.4		
8	1	0.85	1	1	0	0	0	0.5		
9	0	0.25	0	0	0	0	1	0.6		
10	1	0.75	1	1	0	0	0	0.7		
11	0	0.55	1	0	1	0	0	0.8		
12	0	0.75	1	0	1	0	0	0.9		
13				4	2	1	3	1	0	0
14										
15	TP	FP		4	2					
16	FN	TN		1	3	Threshold:	0.5			
17						FPR:	0.4			
18						TPR:	0.8			

How to make an ROC (13)



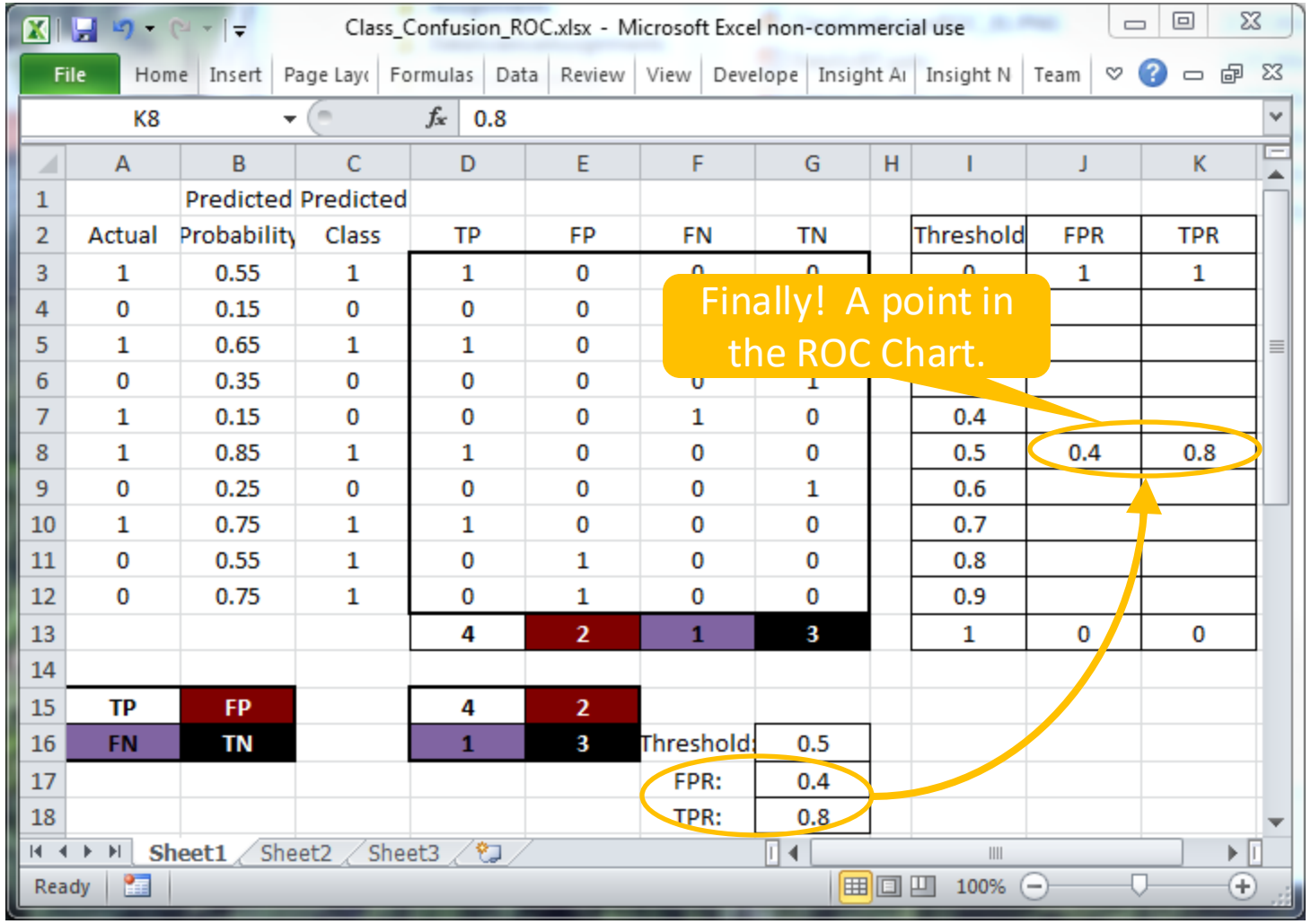
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

How to make an ROC (14)



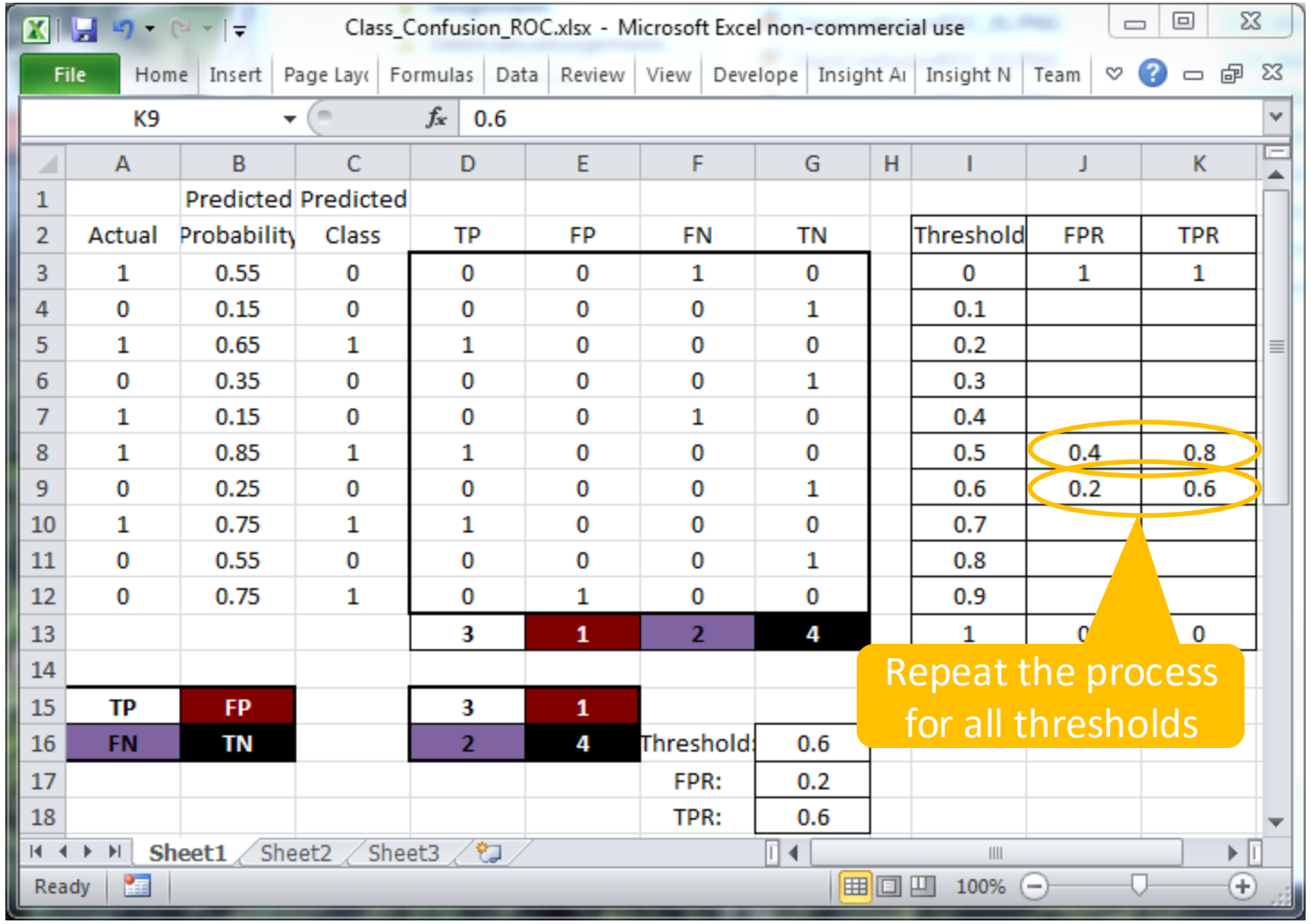
	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5		
9	0	0.25	0	0	0	0	1		0.6		
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	1	0	1	0	0		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				4	2	1	3		1	0	0
14											
15	TP	FP		4	2						
16	FN	TN		1	3	Threshold:	0.5				
17						FPR:	0.4				
18						TPR:	0.8				

How to make an ROC (15)



	A	B	C	D	E	F	G	I	J	K
1		Predicted	Predicted							
2	Actual	Probability	Class	TP	FP	FN	TN	Threshold	FPR	TPR
3	1	0.55	1	1	0	0	0	0	1	1
4	0	0.15	0	0	0					
5	1	0.65	1	1	0					
6	0	0.35	0	0	0	0	1			
7	1	0.15	0	0	0	1	0	0.4		
8	1	0.85	1	1	0	0	0	0.5	0.4	0.8
9	0	0.25	0	0	0	0	1	0.6		
10	1	0.75	1	1	0	0	0	0.7		
11	0	0.55	1	0	1	0	0	0.8		
12	0	0.75	1	0	1	0	0	0.9		
13				4	2	1	3	1	0	0
14										
15	TP	FP		4	2					
16	FN	TN		1	3	Threshold:	0.5			
17						FPR:	0.4			
18						TPR:	0.8			

How to make an ROC (16)

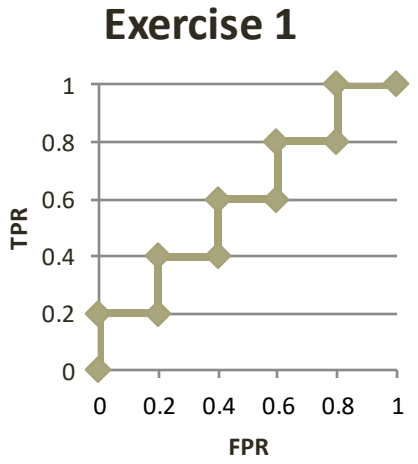


	A	B	C	D	E	F	G	H	I	J	K
1		Predicted	Predicted								
2	Actual	Probability	Class	TP	FP	FN	TN		Threshold	FPR	TPR
3	1	0.55	0	0	0	1	0		0	1	1
4	0	0.15	0	0	0	0	1		0.1		
5	1	0.65	1	1	0	0	0		0.2		
6	0	0.35	0	0	0	0	1		0.3		
7	1	0.15	0	0	0	1	0		0.4		
8	1	0.85	1	1	0	0	0		0.5	0.4	0.8
9	0	0.25	0	0	0	0	1		0.6	0.2	0.6
10	1	0.75	1	1	0	0	0		0.7		
11	0	0.55	0	0	0	0	1		0.8		
12	0	0.75	1	0	1	0	0		0.9		
13				3	1	2	4		1	0	0
14											
15	TP	FP		3	1						
16	FN	TN		2	4	Threshold:	0.6				
17						FPR:	0.2				
18						TPR:	0.6				

How to make an ROC (17)

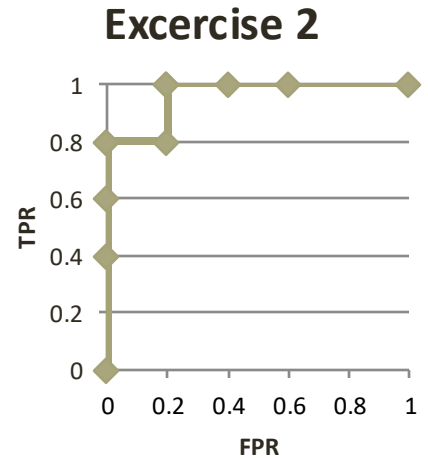
Actual	Predicted Probability
1	0.55
0	0.15
1	0.65
0	0.35
1	0.15
1	0.85
0	0.25
1	0.75
0	0.55
0	0.75

FPR	TPR
1	1
0.8	0.8
0.7	0.7
0.6	0.6
0.4	0.6
0.4	0.6
0.1	0.6
0.1	0.6
0.1	0.6
0.1	0.1
0	0



Actual	Predicted Probability
0	0.15
0	0.25
0	0.35
1	0.45
0	0.45
1	0.55
0	0.65
1	0.75
0	0.85
1	0.95

FPR	TPR
1	1
0.9	0.9
0.9	0.8
0.9	0.8
0.9	0.8
0.1	0.5
0.1	0.5
0.1	0.2
0.1	0.1
0.1	0.1
0	0



How to make an ROC

ROC Misconceptions

- ~~One way to visually evaluate a binary classification model is using the ROC plot. By itself, it is not very useful, but by comparing the ROC plot of multiple models we can start seeing which models are better. The area under the ROC plot is called AUC (area under the curve) and the closer it is to 1, the better the model.~~
- One way to visually evaluate a binary classification model is using the ROC plot. The ROC plot is the only metric that can be used by itself and that has some meaning across different data sets. The area under the ROC plot is called AUC (area under the curve) and the closer it is to 1, the better the model. An AUC of 0.5 is a random model and is considered to be the worst-case scenario. Such a model is represented by the diagonal from the lower left to the upper right. (A model that performs below 0.5 with confidence just needs its labels reversed.)

Review Terminology

- Algorithm
- Anomaly detection
- Association
- Attribute
- Binarize Categories
- Binary Column
- Case
- Category Column
- Character Column
- Classification
- Clustering
- Coercion
- Column
- Column Header
- Confusion Matrix
- Data
- Data Dimensionality
- Data Frame
- Data Type
- DFD
- Dummy Variable
- Estimation
- False Positive
- False Negative
- Feature Scaling
- Field
- Hypothesis
- Key Column
- Machine Learning
- Matrix
- Missing Data
- Model
- Multinomial Column
- Normalization
- Numeric Column
- Observation
- One-hot encoding
- Outcome
- Outlier Removal
- Predictive Analytics
- Rectangular Data
- Relabeling
- ROC curve
- Row
- Schema
- Shaping Data
- Sparse Multi-Dimensional Matrix
- Standard Deviation
- States
- String
- Supervised Learning
- Support
- Table
- Target Column
- Text Column
- Theory
- Un-structured Data
- Unsupervised Learning
- Z-score

Introduction to Data Science