Model Evaluation

Myra O'Regan

October 17, 2017

Model Evaluation

- Focussing on classification models with 2 classes
- Logistic regression
- Support Vector machines
- Classification Trees
- Ensembles

First Things First

- Look at model output
- What would we look at in Trees?
- Big question: Does it make sense?

Test vs Training set

- Split data into Training and Test (80 vs 20)
- Good to have same % of cases in each section
- Use Stratified sampling
- Model built on Training
- Model evaluated on Test set
- Set test should reflect future data !!!

What have we got to work with?

- Test data set - independent and target variable for each case
- Using model we can calculate predicted probabilities p_i of belonging to each class
- Remember $p_1 + p_2 = 1$
- Set of corresponding target values
- What is the question?

Two questions

- Are we trying to evaluate how close the predicted probabilities are to the true probabilities?
- Are we trying to see if this model is good for classification

Two models

Two Models

Model 1		Model 2	
p(1)	Target value	p(1)	Target value
0.95	1	0.24	1
0.94	1	0.24	1
0.93	1	0.22	1
0.85	1	0.22	1
0.20	0	0.20	0
0.15	0	0.10	0
0.05	0	0.05	0

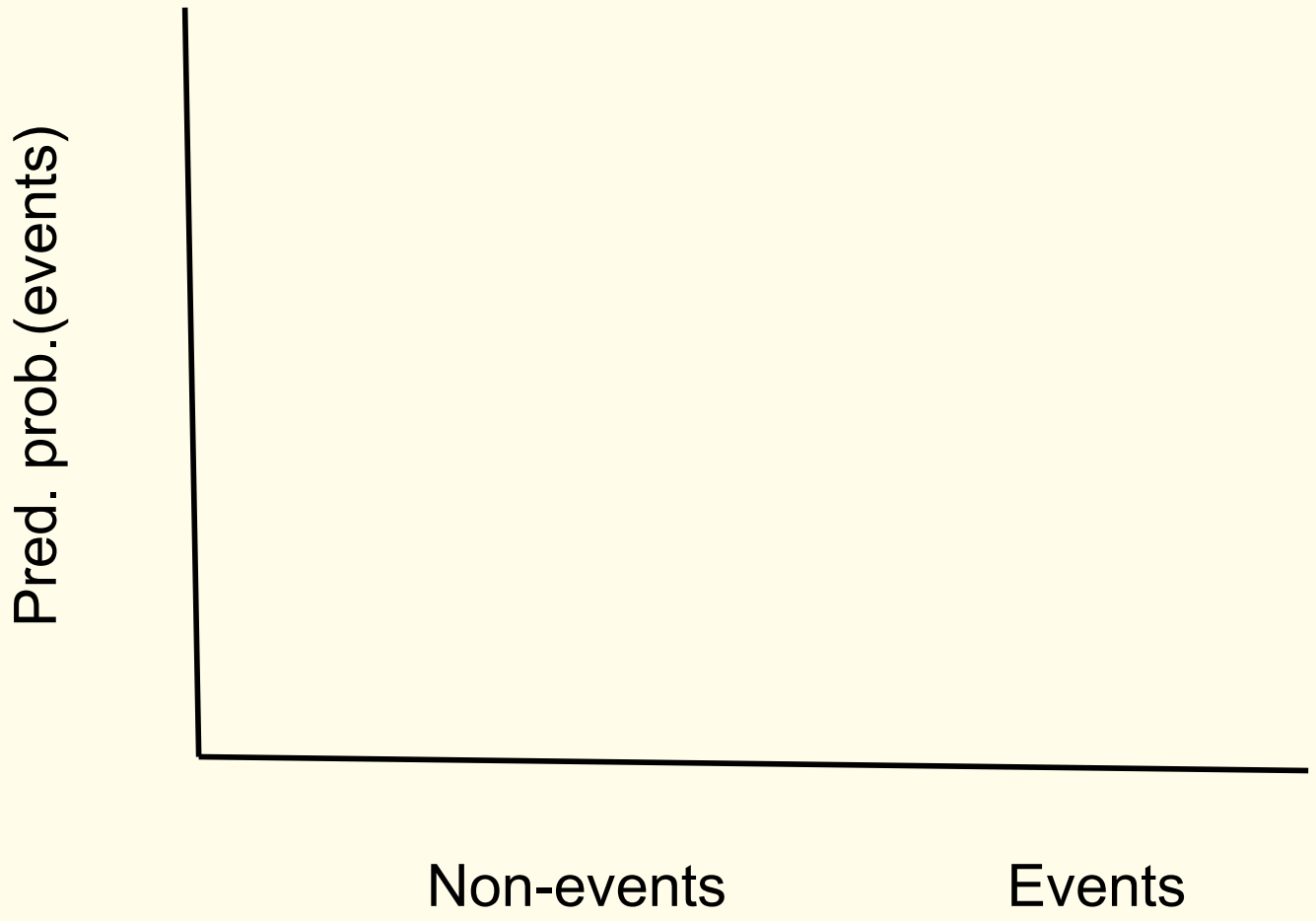
Some Notation

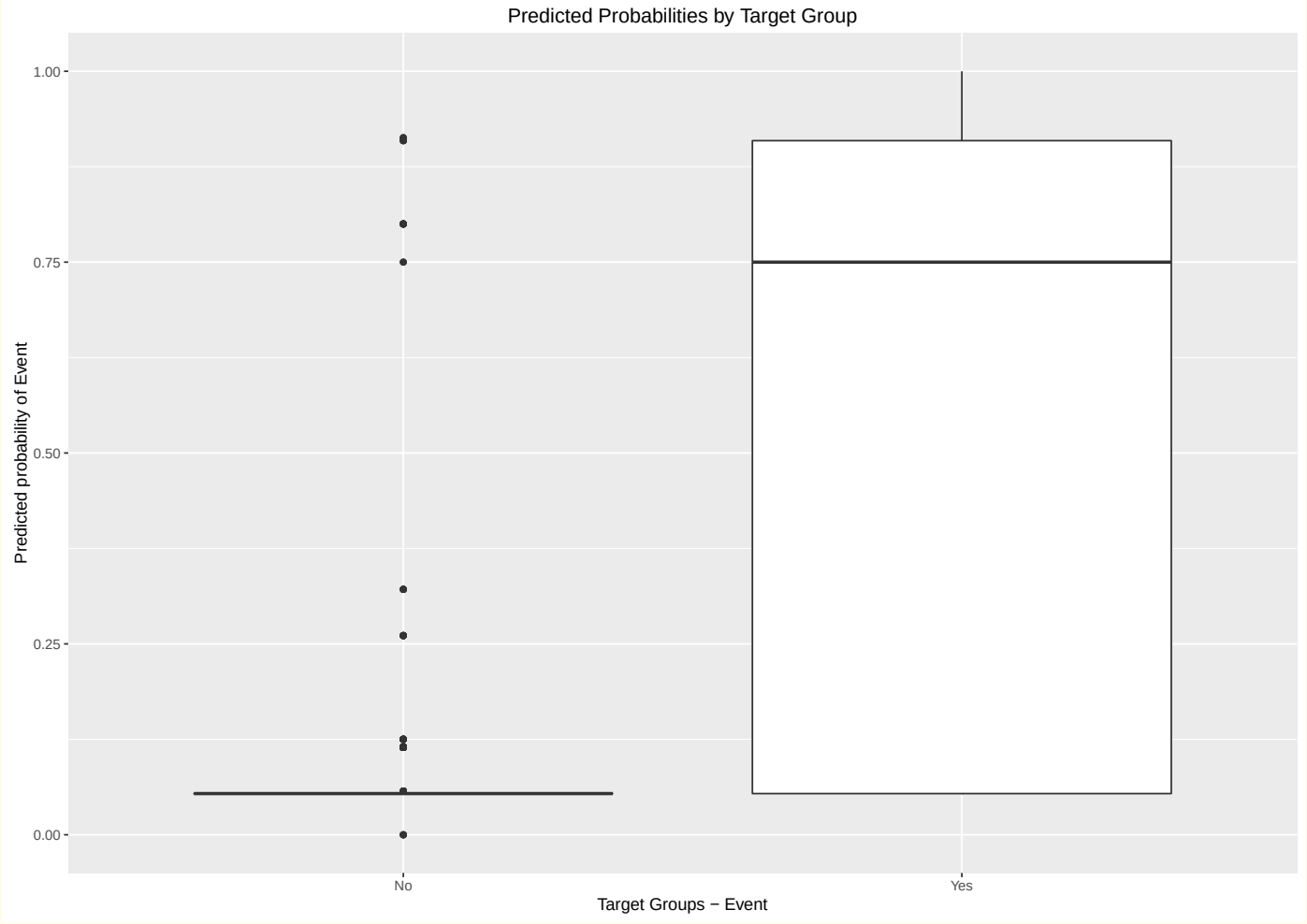
- Events vs Non events
- 1 vs 0
- $+$ vs -

Predicted prob. of event and target values

- Two columns of data
- Sort by predicted prob. of an event
- What would they look like for a good model?
- For a bad model??
- For a random model??
- How could we display the data?

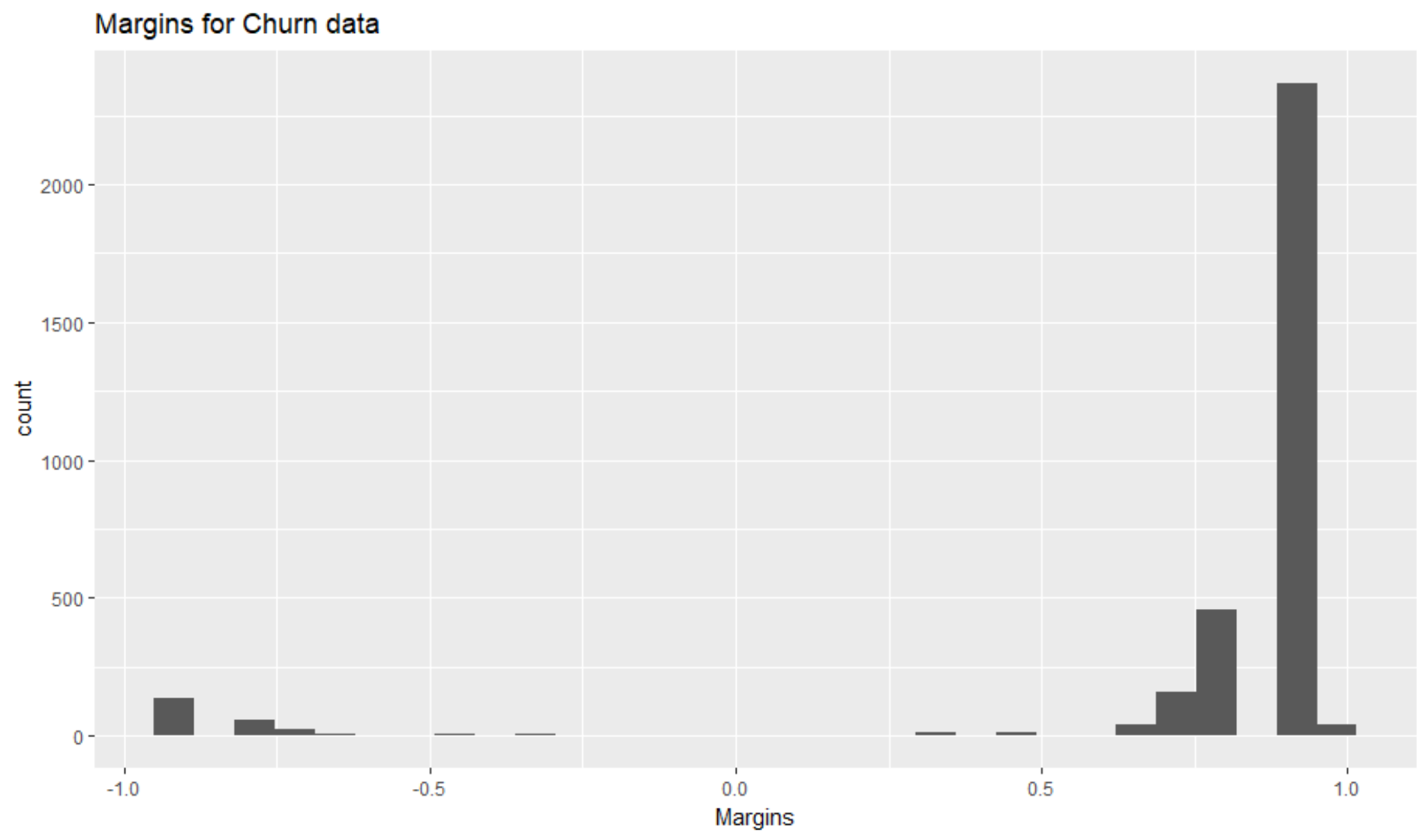
Boxplot

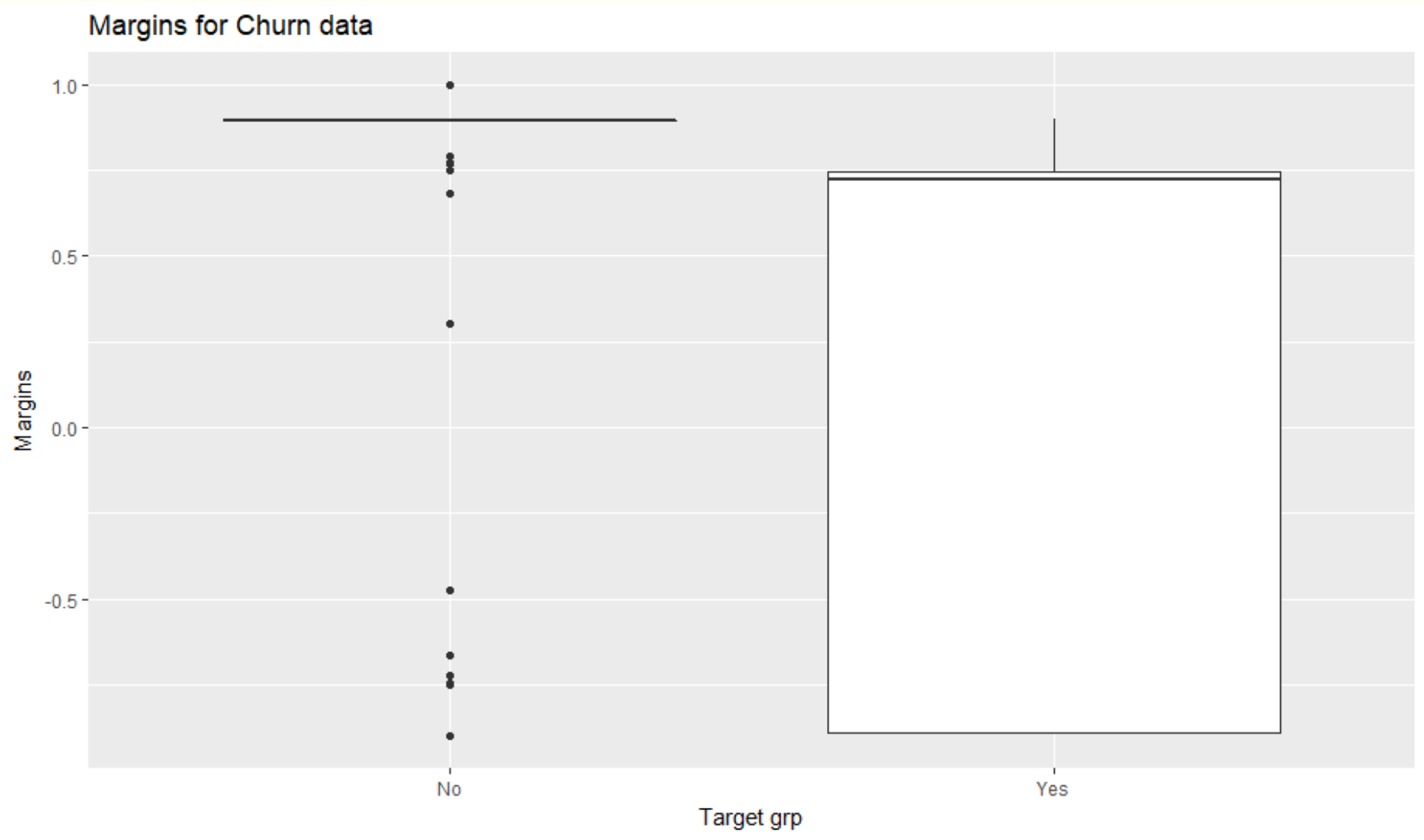



Predicted Probabilities by Target Group
1.00
0.75
0.50
0.25
0.00
Predicted probability of Event
No
Yes
Target Groups − Event

Margins of Classifier

- Predicted probabilities p_1 and p_0 for each of the two target groups
- For a case if target $= 1$ margin $= p_1$-p_0
- If target $=0$ margin $= p_0$-p_1
- Two things here sign and size
- Should these be big or small?


Margins for Churn data
count
2000
1500
1000
500
0
-1.0
-0.5
0.0
0.5
1.0
Margins


Margins for Churn data
1.0
0.5
0.0
-0.5
Margins
No
Yes
Target grp

Confusion Tables

- Pick a threshold c
- For each predicted value of an event pp_i
- If $pp_i <= c$ predict 0 or - (non-event)
- If $pp_i > c$ predict 1 or + (event)

Confusion Table

	Predicted	
Actual	+	-
+	TP	FN
-	FP	TN

TP: True Positives; FN = False Negatives

FP: False Positives; TN= True Negatives

Summary Stats for Confusion tables

	Predicted	
Actual	+	-
+	TP	FN
-	FP	TN

$$\text{Accuracy} = 100 * \frac{TP+TN}{TP+FN+FP+TN}$$

Misclassification error = 100-Accuracy

Summary Stats for Confusion tables

	Predicted	
Actual	+	-
+	TP	FN
-	FP	TN

Measure of accuracy for predicting target events (+)

$$\text{Sensitivity} = 100 * \frac{TP}{TP+FN}$$

Summary Stats for Confusion tables

	Predicted	
Actual	+	-
+	TP	FN
-	FP	TN

Measure of accuracy for predicting non-target events (-)

$$\text{Specificity} = 100 * \frac{TN}{FP+TN}$$

Summary Stats for Confusion tables

	Predicted	
Actual	+	-
+	200	300
-	100	400

$$\text{Senstivity(TPR)}=100 * \frac{200}{200+300} = 40\%$$

$$\text{Specificity(TNR)} = 100 * \frac{400}{400+100} = 80\%$$

$$\text{Accuracy} = 100 * \frac{200+400}{200+400+300+100} = 60\%$$

Example 1

	Predicted	
Actual	+	-
+	400	100
-	300	200

Accuracy = 60%
Sensitivity = 80%
Specificity=40%

Example 2

	Predicted	
Actual	+	-
+	200	300
-	100	400

Accuracy = 60%
Sensitivity = 40%
Specificity=80%

Example 3

	Predicted	
Actual	+	-
+	2500	2500
-	10000	85000

Accuracy = 88%

Sensitivity = 50%

Specificity=89%

Example 3 again

5,000 events denoted by $+$
95,000 non-events denoted by -
Suppose we had no model what would we do?
Assign all to -
No Information rate $= 0.95$

Another example

- A model was built to detect landmines
- Let event = land mine present
- Sensitivity vs Specificity vs Accuracy here

Some other stats

	Predicted	
Actual	+	-
+	A	B
-	C	D

Pos pred. value=A/(A+C);
Neg pred. value= B/(B+D);
Prevalence =(A+B)/(A+B+C+D)

Some other stats

	Predicted	
Actual	+	-
+	A	B
-	C	D

Detection rate =A/(A+B+C+D);
Detection Prev. = (A+C)/(A+B+C+D)
Balanced accuracy=(Sensitivity + Specificity)/2

Mc Nemar test

	Predicted	
Actual	+	-
+	A	B
-	C	D

Look at off diagonal cells

If no difference we would expect $\frac{B+C}{2}$ in each cell

Construct a chi-square to compare the expected value to observed value

Calculation of Mc Nemar

$$\chi^2 = \frac{(Obs-Exp)^2}{Exp} = \frac{(B-\frac{B+C}{2})^2}{\frac{B+C}{2}} + \frac{(C-\frac{B+C}{2})^2}{\frac{B+C}{2}}$$

This reduces to $\frac{(|B-C|-1)^2}{B+C}$

df =1

Another use of Mc Nemar

		A misclassified	
		Yes	No
B Misclassified	Yes	A	B
	No	C	D

Kappa Statistic

		Predicted		
		+	-	Total
Actual	+	100	100	200
	-	300	500	800
		400	600	

Kappa Statistic

- Sometimes called Cohen's Kappa Statistic
- Agreement between two raters
- Look at agreement taking into account the accuracy that would be generated by chance
- Printed by the **caret** package
- Measures relative improvement over random predictor

Calculations

$$Kappa = \frac{(O-E)}{(1-E)}$$

- O is the observed accuracy
- E is the accuracy expected by chance
- Values range form -1 to 1
- 0 no agreement
- 1 is perfect agreement

And more calculations ....

		Predicted		
		+	-	Total
Actual	+	100	100	200
	-	300	500	800
		400	600	1000

Observed accuracy = (100+500)/1000=0.6

What accuracy would you expect by chance?

		Predicted		
		+	-	Total
Actual	+	80		200
	-		480	800
		400	600	1000

Expected accuracy = (80+480)/1000=0.56

Kappa = (0.6-0.56)/(1-0.56)=0.09

Another Example

		Predicted		
		+	-	Total
Actual	+	1900	100	2000
	-	2800	79200	82000

E=0.943; O=0.988; Kappa = 0.786

Measure of uncertainty

Another Example

		Predicted		
		+	-	Total
Actual	+	1500	500	2000
	-	16000	66000	82000

E=0.79; O=0.817; Kappa = 0.129

ROC curve

- Receiver Operating Curve
- Plot Sensitivity -True positive rate (TPR)
- vs
- (100-Specificity) -False Positive rate (FPR)
- For all cutpoints

Very simple example - ROC curve

P(event)	Target Value
0.8	1
0.8	1
0.8	0
0.8	0
0.35	1
0.35	0
0.12	1
0.12	0
0.12	0

Classification tables

Cutoff=0

		Predicted		
		+	-	Total
Actual	+	4	0	4
	-	6	0	6

Sensitivity(TPR)=1 or 100%;
Specificity(TNR)= 0;
False positive (FPR)= 1 or 100%